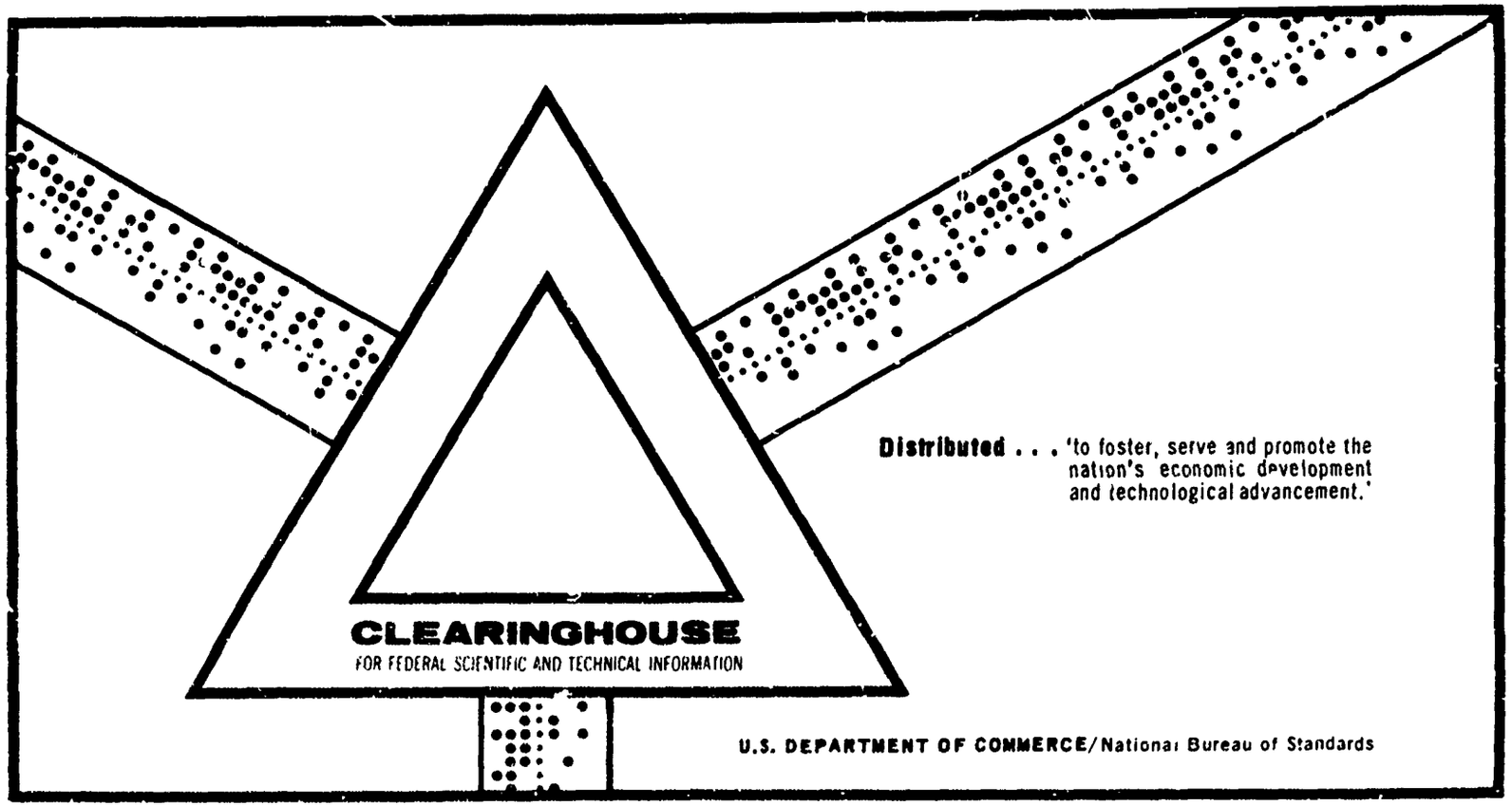
ECONOMICS OF INFORMATION SYSTEMS

Jacob Marschak

California University
Los Angeles, California

November 1969



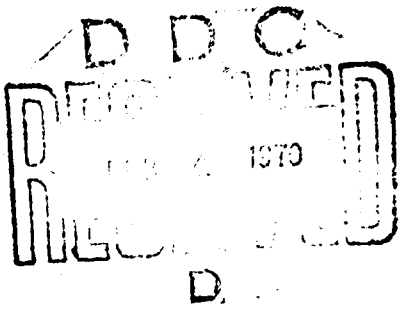
Working Paper No. 153

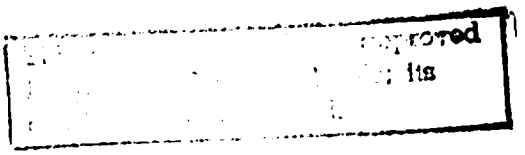
ECONOMICS OF INFORMATION SYSTEMS

by

JACOB MARSCHAK

November, 1969





WESTERN MANAGEMENT SCIENCE INSTITUTE

University of California, Los Angeles

Working Paper No. 153

ECONOMICS OF INFORMATION SYSTEMS

by

JACOB MARSCHAK

November, 1969

This research was supported partially by the Office of Naval
Research under Contract NC0014-69-A-0200-4005, NR 047-041,
and by the National Science Foundation under Grant GS 2041

Western Management Science Institute
University of California, Los Angeles

ECONOMICS OF INFORMATION SYSTEMS by Jacob Marschak

Econometric Society, New York, 29 December 1969

CONTENTS

4. Comparative Informativeness

4.1 Definition.

4.2 Garbling.

4.3 Maximal and minimal information matrices.

4.4 Comparative coarseness.

4.5 Blackwell's Theorem.

4.6 The case of noiseless information.

4.7 Strong ordering by informativeness.

5. Informativeness of Systems over Time

5.1 Environment, action, and observation as time-sequences.

5.2 Effect of memory length on informativeness.

5.3 Delayed vs. prompt perfect information.

5.4 Perfect information with long vs. short delay when the environment is Markovian.

5.5 Obsolescence and impatience.

5.6 Sequential inquiries and adaptive programming.

6. Optimal Inquiries

6.1 Binary information matrices as an example.

6.2 Informativeness of binary inquiries.

6.3 Symmetric binary information matrices.

6.4 Benefit matrix and information value: the case of two actions.

6.5 The case of more than two actions.

6.6 Cost conditions.

6.7 Cost linear in channel capacity.

6.8 Cost of inferring sign of mean of finite population from sign of mean of sample.

7. Economics of Communication

7.1 The fidelity criterion as benefit.

7.2 Capacity of noiseless channel.

7.3 Minimum expected length of code word, as the "uncertainty at source".

7.4 Noisy channel: transmission rate and capacity.

7.5 Capacity and cost.

7.6 Does informativeness always increase with "information transmitted?"

ABSTRACT

An information system is a chain (or, more generally, a network) of symbol-processing components, each characterized by costs and delays, and by the probabilities of its outputs, given an input. In recent times, statisticians, engineers, and even philosophers have all shown increasing tendency to accept the economist's way of comparing information systems according to their average costs and benefits,--the former depending, in part, on the delays between the events inquired about and the actions decided upon.

Statisticians have concentrated on the economic choice of only these two, the initial and the terminal components of the system: "inquiry" and "decision rule". And they have tended to neglect the processing delays arising in these as well as in the intermediate components of a system. Engineers, on the other hand, have concentrated on the intermediate components that form the "communication sub-chain": "memorizing", "encoding", "transmitting", "decoding". And they have been concerned with the processing delays that depend on the average number of code symbols needed (and thus on the "entropy" to be removed by communication).

For simplicity, we have assumed that utility (the quantity whose expected value is maximized by the user) is the difference between costs and benefits. The current literature on communication assumes implicitly that other choice criteria

(such as the length of a code word) are also additive, and that channels with equal capacity are equally costly. These assumptions may need to be qualified, by studying channel costs and the economic effects of communication delays.

The economically minded user must consider the several system components jointly; and it turns out that, in certain important cases, the average difference between the benefit and cost to a user is maximized by large-scale demand. Moreover, the aggregate demand of all users will depend on the joint supply conditions for the various system components. It will thus depend, for example, on the cost economies due to the "packaging" of several components, to standardization and large-scale production. This opens up the question whether social interest is best served by a competitive market in information processing equipment and services, human as well as inanimate.

0. Introduction

0.1 The economist's general information problem. Out of several pushbuttons, each of a different color, you select one. A slight push, and massive amounts of energy are released, and are transformed in the manner you have prescribed. The button colors which you have perceived and from which you have selected, exemplify signs, symbols. Your "manipulation of symbols", equally vaguely called "handling of information" has involved little energy but has discharged and directed a large amount. You have done "brain work." No economist will deny that a large part of our national product is contributed by symbol manipulation -- telephoning orders, discussing in conferences, shuffling papers, or just performing some of the humble tasks required of the inspector, or even an ordinary worker, on the assembly line.*)

*) See Marschak [1968A], a paper addressed to a wider audience and, in essence, revised here in a somewhat more precise fashion. For some earlier results see Marschak [1954]. Much is owed to discussions with J. MacQueen. END OF FOOTNOTE.

The economist asks, first: what determines the demand and supply of the goods and services used to manipulate symbols. This may help him, second, to understand how social welfare is affected by the manner in which resources are allocated to those goods and services.

A pre-requisite is, to define concepts and study their interrelations in a way that would prove useful for the answering of these questions. The economist begins by assuming that those who demand and use, and those who produce and supply, the goods and services considered, make choices that are "economical" (= "rational") in some usefully defined way, and are made under well-defined constraints. The constraints may include limitations on the choosers' memories and other abilities. The economic theorist leave the door open to psychologists, sociologists, historians, and to his own 'institutionalist" colleagues in the hope they will help to determine the values of underlying parameters,--provided (another hope!) they do not establish that the assumption of "economical" choice fails to yield usefully close approximations to begin with. I take this back: even then, he will offer his results as recommendations to users and producers of "information-handling", or "informational", goods and services.

0.2 <u>The user's problem, viewed by non-economists</u>. Besides its interest to economists, the manipulation of symbols, or information processing, has been the domain of philosophers and linguists; of computer scientists, control theorists and communication engineers; and of statisticians. The latter, following the path of J. Neymann and A. Wald, have become

more and more concerned with the economical manner of obtaining "information", and have discovered much that is useful to the economist. Engineers have proposed a measure of "information" / "transmitted" based on probability relations between classes of arbitrary signs. This arose out of practical, "economic" needs of the communication industry. My task will be, in part, to see how those results fit into the general economics of symbol-manipulating goods and services,--including, for example, the services of statisticians, and of men who design or handle computers and control mechanisms. (The task of the last-named men is indeed to apply economics to to-day's most varied and complete combinations of informational goods and services!).--Finally, attempts have been made on the part of philosophers and linguists*) to modify the engineers'

*) See e.g., Carnap and Bar-Hillel [1952]; somewhat differently, Miller and Chomsky [1963]. END OF FOOTNOTE

measure into 'semantic information" or "content" measure--essentially by substituting for a class of arbitrary signs its partition into equivalence classes consisting of signs with identical "content" ("meaning").

In recent years, the approach <u>via</u> economic rationality--(bluntly: via the expected utility to the decision maker)--has begun to penetrate the work of both engineers and philosophers. An important, though still not sufficiently well known step,

was made by pioneer C. Shannon himself [1960] when he removed his earlier tacit assignment of equal penalty for all communication errors. He introduced, instead, a "fidelity criterion". This is indeed utility itself--albeit confined (as we shall see) to the context of communication only and therefore defined on a very special class of actions and events. And Ronald A. Howard [1966] writes, in a broader context:

> "...The early developers stressed that the information measure was dependent only on the probabilistic structure of the communication process. For example, if losing all your assets in the stock market and having whale steak for dinner have the same probability, then the information associated with the occurence of either event is the same. ...No theory that involves just the probabilities of outcomes without considering their consequences could possibly be adequate in describing the importance of uncertainty to the decision maker."

He concludes his analysis of a neat model with a challenge to his profession (and perhaps to mine as well):

> "If information value and associated decision theoretic structures do not in the future occupy a large part of the education of engineers, then the engineering profession will find that its traditional role of managing scientific and economic resources for the benefit of man has been forfeited to another profession."

And philosopher R. Carnap whom we have mentioned as one of the early proponents of a "semantic" information measure ("content measure') wrote in a more recent [1966] paper:

> "When I consider the application of the concept of probability in science then I usually have in mind in the first place the probability of laws or theories. Once we see clearly which features of prediction are desirable, then we may say that a given theory is preferable to another one if the predictions yielded by the first theory possess on the average more of the desirable features than the prediction yielded by the other theory."...

He then proceeds to show that if 'a practically acting man"

> "bases his choice either on content measure alone or on probability alone, he will sometimes be led to choices that are clearly wrong." "We should choose that action for which the expectation value of the utility of outcome is a maximum."(pp. 252, 253-4, 257).*)

---

*) In the quoted paper, he also says that another paper [Carnap, 1962] (strongly influenced by Ramsey, De Finetti, and Savage) "gives an exposition of my view on the nature of inductive logic which is clearer and from my present point of view more adequate than that which I gave in my book," viz. in Carnap [1950].

0.3 Individual demand for information services. Thus encouraged by the spread of understanding of the economic approach to information use, I shall proceed with my task, a more special one than the general economic information problem outlined at the beginning. I shall study the rational choice-making of an individual from among available information systems, or available components of such systems. The availability constraint specifies, in particular, the costs and the delays associated with given components, or with chains or networks of components, of information systems. As is familiar to students of the market, the available set depends on the choices made by suppliers. In last effect, joint choices by demanders and suppliers would determine which information systems are in fact produced and used under given external conditions. These conditions include the technological knowledge of those concerned.

I shall not be able to make more than casual remarks on the supply The first of the two general questions to be asked by the economist, the joint determination of demand and supply, will therefore receive only a partial answer . The second question, that of socially optimal allocation of resources to informational goods and services, is pushed away still farther. This is not to say that the allocation question cannot be studied till the demand and supply of informational goods and services is fully understood. Significant work of Hurwicz [1960], Stigler [1961,1962], Hirshleifer [1967], Radner [1967, 1968] testifies to the contrary.

1. PROCESSING

1.1 Processing P is defined as

$$P = \langle X, Y, \eta, \gamma, \tau \rangle , \text{ where}$$

X = set of inputs x

Y = set of outputs y

η = transformation from X to Y, including the case of stochastic transformation (see below)

γ = transformation from X to non-negative reals, measuring cost (in cost units)

τ = transformation from X to non-negative reals, measuring delay (in time units)

X, Y are, generally, random sets. As to η: in a special case called "deterministic" or "noiseless," η is an ordinary function; i.e., it associates every x in X with a unique $y = \eta(x)$ in Y. However, we must consider the more general case, called "stochastic" or "noisy," in which, instead, η associates every x in X with some ("conditional") probability distribution on Y. For simplicity of presentation we shall usually (except for some economically interesting examples) assume X and Y finite,

$$X = (1,\ldots,m), \quad Y = (1,\ldots,n) ,$$

so that $\eta_{ij} = \text{Prob}(y=j \mid x=i)$. Hence $\eta = [\eta_{xy}]$ is a $m \times n$ Markov matrix, i.e., all $\eta_{xy} \geq 0$ and $\sum_y \eta_{xy} = 1$ for all x. But see Blackwell [1953] for an extension of the concept of stochastic transformation to

infinite sets. Clearly, the special, deterministic case occurs if one element in each row of the matrix $[\eta_{xy}]$ is = 1; then we can write

(1.1.1) $$\eta_{xy} = \begin{matrix} 1 \\ 0 \end{matrix} \text{ if } y \begin{matrix} = \\ \neq \end{matrix} \tau(x) .$$

As to γ: we shall assume $\gamma(x)$, the cost of processing a given input $\underline{x}$, to be constant. We thus forego the discussion of a more general, stochastic case, in which $\gamma(x)$ is a probability distribution of costs, given $\underline{x}$. Similarly, we assume that the time $\tau(x)$ required to process a given input $\underline{x}$ is constant.

1.2 Cost-relevant-inputs. In important cases, exemplified by processings called "storage" and "transportation," two otherwise different inputs, $x = i$ and $x = i'$, say, are such that $\gamma(i) = \gamma(i')$. (It costs the same to transport, over 100 miles, a gallon of whiskey or of gasoline.) It is then convenient to replace the original set $\underline{X}$ by a reduced set X/γ consisting of equivalence classes x/γ such that all elements of the same class are associated with the same cost.

1.3 Available (feasible) processings. For given $\underline{X}$, $\underline{Y}$, not all triples (η, γ, τ) are available. For example, to implement a given transformation η at lowered delays $\tau(x)$ for all $\underline{x}$ may require raised costs $\gamma(x)$. The set of available processings will be denoted by ρ.

1.4 Purposive processing. Consider a case in which the $\underline{y}$ in $\underline{Y}$ [now to be rewritten as $\underline{a}$ in $\underline{A} = (1,\ldots,n)$] can be interpreted as the actions (decisions) of a person who obeys certain axioms of decision

logic[1], and the inputs $\underset{\text{in } X}{\underline{x}}$ / [now to be rewritten as $Z = (1,\ldots,m)$] are events beyond his control. Then there exists a probability distribution π = vector $[\pi_z]$ and a bounded real-valued "utility function" $v(a,z,\gamma(z),\tau(z))$ such that, given two available processings

$$P' = \langle Z',A',\pi',\gamma',\tau' \rangle; \quad P'' = \langle Z'',A'',\pi'',\gamma'',\tau'' \rangle ,$$

the chooser of a processing will choose P' only if

$$U_{\pi v}(P') \geq U_{\pi v}(P'') ,$$

where, for any processing $\underline{P}$, its (expected) utility is

$$(1.4.1) \qquad U_{\pi v}(P) = \sum_z \sum_a \pi_z \gamma_{za} v(a,z,\gamma(z),\tau(z)) .$$

It follows that, given the characteristics of the chooser (viz., π, v, listed in the subscript under U for convenience) and given the available set $\mathcal{P}$, processing $\underline{P}^*$ will be chosen only if

$$P^* \in \mathcal{P}, \quad U_{\pi v}(P^*) \geq U_{\pi v}(P), \quad \text{all } P \text{ in } \mathcal{P} .$$

Note that "chooser" was the word used, instead of "decision-maker": see also Sec. 2.2 where the chooser of $\underline{P}$ will be called meta-decider.

1.5 Timing. Utility depends on action. Accordingly, we consider that the utility is "earned,"

1/ I refer to the work of F.P. Ramsey, B. De Finetti, L.J. Savage, accepted in recent years by professional logicians R. Carnap and R.C. Jeffrey. For a survey see Marschak [1968B]; also, regarding Carnap and regarding the relation of probability to frequency see Marschak [1970]. That certain observed behavior is not really inconsistent with the expected utility rule if cost or feasibility of storing or other processing is accounted for, was brillantly shown by S. Winter [1966]. Among the many merits of Raiffa's delightful introduction to the field [1968] is his forceful emphasis on the need for and the possibility of training people for consistency.

and the action a is taken, at the same time. But the cost $\gamma(z)$ is incurred $\tau(z)$ time units earlier.

1.6 Continued purposive processing. It is often necessary to reinterpret the output a and input z as time-sequences, with "horizon" $= T$, possibly infinite:

$$(1.6.1) \qquad a = \{a_t\},\ z = \{z_t\},\ t = 1,\ldots,T\ .$$

An element η_{za} of the transformation η is then the conditional probability of a particular sequence of T actions, given a sequence of T events. Applying the results of Koopmans [1960], the utility $w(a,z,\gamma(z),\tau(z))$ entering the definition (1.4.1) of the utility of processing can be decomposed thus:

$$(1.6.2) \qquad w(a,z,\gamma(z),\tau(z)) = \sum_{t=1}^{T} u(\bar{a}_t,\bar{z}_t,\gamma(z_t))d^{\sum_{s=1}^{t}\tau(z_s)}\ ,$$

where the "discount constant" d $(0 < d \leq 1)$ and the function u are independent of time and $\bar{a}_t$, $\bar{z}_t$ are, respectively, the "histories up to t":

$$(1.6.3) \qquad \bar{a}_t = (a_1,\ldots,a_t),\ \bar{z}_t = (z_1,\ldots,z_t)\ .$$

1.7 Additive costs and discounted benefits. A convenient though rather special assumption is often tacitly made in practice. It is assumed that, given any distribution π, the utility of processing, $U_{\pi\eta}(P)$ increases in the "expected discounted benefit," B, and decreases in "expected cost," C. Before defining B and C precisely,

let us state the assumption in two other, obviously equivalent, forms: for any given π, (1) "of all processings with the same $\underline{C}$, the one with highest $\underline{B}$ has highest utility"; and (2) "the efficient subset of ρ consists of all those available processings for which the pair $(-C,B)$ is not dominated by any other such available pair."

If, on the other hand, the assumption does not hold, then a processing may exceed in utility, and hence should be chosen in preference to, another processing, even though the latter has lower expected cost and higher expected discounted benefit. It will be shown that the stated tacit assumption implies that the utility function π is decomposable in a certain sense.

More precisely, we define

$$C \equiv C_{\pi}(\gamma) \equiv \sum_{z} \pi_z \gamma(z) , \tag{1.7.1}$$

$$B \equiv B_{\pi\beta d}(\gamma,\tau) \equiv \sum_{z} \sum_{y} \beta(a,z) d^{\tau(z)} {}_{z}\eta_{zy} , \tag{1.7.2}$$

where β is the "benefit function" from $Z \times A$ to reals, and $\underline{d}$ is the discount constant. [Similar to (1.4.1), subscripts under $\underline{C}$, $\underline{B}$ convey the relevant characteristics of the decision maker.] Note that $\underline{d}$ occurs in (1.7.2) but not in (1.7.1). This is because of the assumption on timing in Section 1.5; this difference will be removed when we study processing chains, as in Section 1.9.

It follows from the general theorem on multi-criterion decisions (Appendix I) that $U_{\pi\beta}$ is monotone increasing in $\underline{B}$ and in $-\underline{C}$ if

and only if there exists a function β and constants $\underline{d}$ and $\underline{k}$ such that

(1.7.3) $\omega(a,z,\gamma(z),\tau(z)) = -k\gamma(z) + \beta(a,z)d^{\tau(z)}$; $k > 0$;

($\underline{k}$ is a conversion factor, fixing the choice of units). It then follows by (1.4.1) that the utility of processing is monotone in $-C$ and B (for all π), if and only if it is a linear combination,

(1.7.4) $U = -kC + B$, $k > 0$.

(Elsewhere, $\underline{B}$ was called "expected gross payoff": see Marschak and Radner [in press]).

1.8 Benefit-relevant events and actions. It is convenient to define $\underline{Z}$ and $\underline{A}$ in such a manner that

$$\beta(a,z) = \beta(a,z'), \text{ all } a \in A, \text{ only if } z = z' ,$$

and $$\beta(a,z) = \beta(a',z), \text{ all } z \in Z, \text{ only if } a = a' .$$

Thus if Z and A are finite, so that β can be represented as a "benefit matrix"

$$\beta = [\beta_{az}] ,$$

no two columns and no two rows are identical. (Returning to Section 1.2, we note that z and z' may be equivalent with respect to costs but not equivalent with respect to benefits.) And no generality is lost if all the dominated rows are deleted.

1.9 Processing chains. Define a sequence

$$P^1,\ldots,P^K, \text{ where}$$

(1.9.1) $$P^k = \langle X^k, X^{k+1}, \eta^k, \gamma^k, \tau^k \rangle , \quad k = 1, \ldots, K .$$

Let X^k have n_k elements, so that η^k is of order $n_k \times n_{k+1}$. Such a sequence is equivalent to a processing

(1.9.2) $$P = \langle X^1, X^{K+1}, \eta, \cdot , \cdot \rangle ,$$

where

$$\eta_{x^1 x^{K+1}} = \sum_{x^2} \cdots \sum_{x^K} \eta_{x^1 x^2} \eta_{x^2 x^3} \cdots \eta_{x^K x^{K+1}}, \quad \text{so that}$$

(1.9.3) $$\eta = \prod_{k=1}^{K} \eta^k .$$

With P in (1.9.2) equivalent to the sequence (1.9.1), the values achieved by P and by that sequence should be, in a purposive case, equal. This makes it impossible, in general, to fill the places indicated by dots in (1.9.2) by single real-valued functions. Rather, the utility of $\underline{P}$ (if $\underline{P}$ is purposive) would depend on the sequences $\{\gamma^k\}$, $\{\tau^k\}$, $k = 1, \ldots, K$. This is easily seen by applying the decomposition of utility over time as in (1.6.2) to the case (1.7.3) of additive costs and benefits.

1.10 Networks. More general than a chain is a network, in which each transformation may have several input and output variables, some possibly shared with other transformations. We shall not pursue this here. See Marschak and Radner [in press], Chapter 8.

2. SYMBOLS AS OUTPUTS AND INPUTS

2.1 A purposive processing chain. Consider a chain (1.9.1) consisting of K successive processing links, with

$$X^{K+1} = \text{set } \underline{A} \text{ of actions } \underline{a} ,$$

$$X^{1} = \text{set } \underline{Z} \text{ of events } \underline{z} ,$$

where $\underline{a}$ and $\underline{z}$ are typical arguments of the benefit function $\beta(a,z)$. Each may be a time-sequence as in (1.6.1). Some physical processes cause an action and event to jointly yield some physical consequence (again possibly a time-sequence: e.g., a sequence of annual monetary profits), to which a benefit number is attached. But we shall not be concerned with these physical processes, and/chains (or networks) that the they form.

The inputs and outputs of the intermediate processing links, $P^2,\ldots,P^K$ do not enter the benefit function. As in Section 1.2, two elements x_1^k, x_2^k of the set X^k, $k = 2,\ldots,K$, can be considered equivalent if their processing costs are equal:

$$\gamma^k(x_1^k) = \gamma^k(x_2^k)$$

It will be convenient to reserve the term "symbols" for these, "benefit-neutral" but "cost- relevant," inputs and outputs. Thus the links $P^2,\ldots,P^K$ will be said to process symbols onto symbols. Typical examples are: translation (e.g., encoding, decoding) of messages; transmission of messages over distances; and their storage over time. On the other hand, an event or an action (even that of a painter or composer)

will not be called a symbol; but processing link P^1 will be said to transform event into a symbol; and P^K will be said to transform a symbol into action.

2.2 Choosing the chain: a meta-decision. The action, or decision, $a \equiv x^{k+1}$, the output of the last link in the purposive chain must be distinguished from the decision to choose one rather than another chain. The difference between expected benefit and cost is maximized by the chooser of the chain. The chooser may hire men or machines to perform the successive processings, including the ultimate one, viz., the choice of action, or decision. If this ultimate processing link is called deciding, the choice of it and of other links of the chain may be called meta-deciding.

2.3 Some information systems. A purposive processing chain is often called an information system, the word information presumably bearing some relation to transformations from and into symbol sets. Information about a physical fact is not the fact itself but some "symbols" (e.g., words) associated with it. Historically, two kinds of "shortened" chains have been considered by specialists: statisticians on the one hand, and communication engineers on the other. They are

(a) a two-link chain, with

$X^1 = Z$ = events — P^1 = experiment, inquiry

$X^2 = Y$ = data, observations — P^2 = strategy

$X^3 = A$ = actions, decisions

(b) a four-link chain, with

$X^1 = Z$ = messages to be sent — P^1 = storing

X^2 = long stored sequences of messages — P^2 = encoding

X^3 = encoded messages P^3 = transmission

X^4 = received messages P^4 = decoding

$X^5 = A =$ = decoded messages

The chains (a) and (b) are linked together on Figure 5.

To suit special applications, some special assumptions are usually made, different in (a) and in (b), regarding the sets of inputs and outputs, the sets of available processings, the cost and delay functions γ and τ, and the benefit function β. We shall indicate some of those assumptions and the implications of removing them, in due course.

Both (a) and (b) can be considered as special cases of some longer chain. It seems that such longer chains are necessary to describe, in their full richness, the operations of a computer (including problem-solving, simulation, pattern recognition, etc.). The popular description of these operations as "information processing" would then appear a felicitous one. This would include, for example, programmed navigation. See Chernoff [1967].

In the following three Sections 3, 4, 5, we deal with the two-links chain (a), and study the consequences of some simplifying assumptions used, in effect, in the literature of "statistical decision theory." These results are, in fact, applicable also to information systems consisting of any number of links, with actions based, not directly on observations (outputs of the "inquiry" link), but on the outputs of subsequent processings (e.g., encoding, transmitting) of observations. By (1.9.3), the system's transformation matrix η is the product of the successive transformation matrices, η^k, of its links; and the latter need not be specified if the assumptions listed in Section 3 are made.

Accordingly, in the next three sections, T will be called, interchangeably, the inquiry matrix or, to be more general, the transformation matrix of an information system or, briefly, information matrix.

2.4 Of the assumptions to be listed in Section 3 that of additive cost is perhaps least offensive and is, at the same time fruitful of important results, for it permits to concentrate on the properties of the information matrix T. On the other hand, the question of successive delays (operation speeds and capacities at successive links), mostly neglected in the two-links theory and introduced in our Section 4 in general terms only, will become a serious one when the processing chain is lengthened by inserting links that implement the communication between the observer and the decision-maker.

3. INQUIRING AND DECIDING IN STATISTICAL THEORY

3.1 The two-link chain. Link P^1 in the two-link chain (a) of Section 2.3 has been variously called "experiment," "taking observations," also "making a diagnosis." Link P^2, "strategy", has been also called 'decision rule." Reflecting certain though surely not all aspects of statistical practice, the usual analysis of the two-link chain makes tacitly some restrictions which do not appear necessary or justifiable in the broader context of economic comparison of purposive processings. In particular, the delays $\tau^1(z)$, $\tau^2(y)$ are neglected; and so are the constraints on strategies, and their cost, $\gamma^2(y)$.

On the other hand, in most statistical writings, our environmental variable z is generalized, as follows. The event (or, in the case of continued processing, a time-sequence of events) is replaced by a probability distribution, called "hypothesis," so that our π becomes a distribution on the space of probability distributions of some variable v. However, this complicated description of the problem is equivalent, and can be reduced, to the original problem, with v playing the role of the event z. We shall, therefore, not pursue this further.

3.2 Neglecting delays. While, as will be shown, the speed of processing is attached great importance in the existing work of communication engineers who study the several-links chain (b) described in Sec. 2.3, processing speed is completely neglected in the statistical theory of the two-links chain (a). No explicit attention is paid to whether it takes an hour or a month to collect a sample, or to apply a given decision rule. Accordingly the question of "overloaded capacity"

of an observation equipment or decision-making equipment is not, to my knowledge, treated explicitly in statistical literature. It is assumed in effect that for all processing chains considered, $\tau^1(z)$ is the same constant, and $\tau^2(y)$ is the same constant; so that, when comparing the values of two processings, one can assume for both, without loss of generality,

$$\tau^1(z) = \tau^2(y) = 0 . \tag{3.2.1}$$

No doubt this assumption is not made in actual statistical practice. If the expected benefit can be strongly diminished when decisions are based on obsolete data (see Section 5), the chooser of the experiment and the strategy will give preference to accelerated ones, costs permitting. Moreover, it is not economical to accelerate the experiment if this results in piling up unused data because decisions are taken too slowly. Such considerations surely arise in industrial quality control, in marketing research, in the preparation of economic indices for public policy, and, very likely also in much of scientific laboratory and clinical work.*/

3.3 With delays out of the way, the "Statistical Decision Problem" takes the following form. Changing notations somewhat, write:

$$\eta^1_{zy} = p(y|z) = \eta_{zy};\ \gamma^1(z) = \gamma_z;\ P^1 = \langle \eta, \gamma \rangle$$
$$\eta^2_{ya} = p(a|y) = \alpha_{ya};\ \delta^1(z) = \delta_z;\ P^2 = \langle \alpha, \delta \rangle . \tag{3.3.1}$$

The sets $\underline{Z}$ and $\underline{A}$ are regarded as fixed; this and the fact that $\underline{Y}$

*/ See, for example, N.G. Anderson [1969].

is the range of η and the domain of α justifies the above abbreviated definition of the links P^1, P^2. Then the processing chain (P^1, P^2), if available, can be written as

$$P \equiv (\eta, \alpha, \gamma, \delta) \in \mathcal{P} , \tag{3.3.2}$$

where $\mathcal{P}$ is the feasible set. We/postpone till later (Sec. 5) the consideration of continued processing introduced in Sec. 1.6. The assume additive cost as in Sec. 1.7 but chooser then, maximizes, subject to the constraint (3.3.2), the expected utility $\underline{U}$ / which is the difference between expected benefit $\underline{B}$ (no discounting for delay need be considered) and expected cost $\underline{C}$, where

$$B = B_{\pi\beta}(P) = \sum_z \sum_y \beta(a,z)\, \pi_z \eta_{zy} \alpha_{ya} , \tag{3.3.3}$$

$$C = C_{\pi}(P) = \sum_z \pi_z \gamma_z + \sum_z \sum_y \pi_z \eta_{zy} \delta_y , \tag{3.3.4}$$

$$U = U_{\pi\beta}(P) = B_{\pi\beta}(P) - C_{\pi}(P) . \tag{3.3.5}$$

As in Section 1.7, the subscripts under $\underline{B}$, $\underline{C}$, $\underline{U}$ characterize the chooser. Together with the feasible set $\mathcal{P}$, they form the givens of the chooser's problem. Hence the optimal chain / P^* is a function of π, β, δ. So is the efficient set, which consists of all elements of $\mathcal{P}$ for which the pair $(-C,B)$ is not dominated by any other such feasible pair.

3.4 Action as a subset of events. In general, there is no need to assume any formal, logical relation between $\underline{Z}$ and $\underline{A}$. For example, $\underline{Z}$ may be the set (cancer, no cancer), and $\underline{A}$ may be the set (surgery, radiotherapy, no treatment). The benefit function β would then assign

a value to each of the 2x3 = 6 pairs (a,z). In statistical literature, an action that can be considered relevant to the benefit of the statistician's "employer," can be identified with the choice of one of disjoint subsets ("alternative hypotheses") of the set of benefit-relevant events. Such actions cannot be more numerous than events.

True, the action of the statistician is, in other cases, said to consist in choosing from a set of overlapping subsets of events: e.g., in naming an interval.*/ He is then supposed to use choice criteria relevant, I think, to his own, not his employer's, benefit. It is difficult to see how, for example, the length of a confidence interval in a market prediction affects the seller's profit, given the state of the market.

For purposes of economics of information, it is more useful to say that the statistician's task is to derive, from observations $\underline{y}$, the likelihoods η_{zy} for all events $\underline{z}$ relevant to his employer's benefit. Given the prior probabilities π_z, one can then determine the joint probabilities $\pi_z\eta_{zy}$ or, for that matter, the posterior probabilities $(\pi_z\eta_{zy}/\Sigma_t\pi_t\eta_{zt})$. The employer or his operations research man (possibly identical with the statistician) will combine these probabilities with the benefits yielded to the employer by his actions, given the events, and choose the action that maximizes expected benefit.

Accordingly, we shall permit the employer's (user's) actions to be more numerous than events. This will lead to interesting results in the economics of comparing information systems: see Section 6.5.

---

*/ Contrast Examples 1-3 with Example 4 in Lehmann [1959], Section 1.2. See also Pratt [1961]. I am indebted to W. Kruskal for discussions of this question. END OF FOOTNOTE.

To be sure, a problem of communication arises. It is, in fact, the problem of optimal encoding, in the sense of our Section 7, below. It may be costly or even non-feasible to communicate in all detail the posterior or the joint probability distributions involved, to the employer, or to his operations research man, or to a low-echelon decision-making man or machine. With this in mind, a condensed message may be used: for example, the posterior probability that $\underline{z}$ lies in a particular interval. The choice of the interval will then depend, not on the statistician's "tastes", but on the "meta-decider's" judgment as to the contributions of alternative codes to his benefit and cost.

3.5 Neglecting the constraints and costs of deciding. In important parts of statistical literature decision-making is, in effect, assumed costless and unconstrained. This strong assumption has led to a fruitful discussion of "comparative informativeness" of the matrices $\pi = [\pi_{zy}]$. We shall pursue it in some detail in Sections 4 and 5.

The assumption of costless and unconstrained deciding is too strong to have been actually accepted in practice. For example, in the case where observations $\underline{y}$ and decisions $\underline{e} = \alpha(\underline{y})$ are both real-valued, attention was paid, quite early, to a special class of decision rules, viz., to the class of linear α, presumably because linear functions require less computational effort. (The theorem that, among unbiased linear estimators the least-squares estimator is best, goes back to Gauss, I understand.) The search for good "robust" statistics is also due to considerations of computational economy, I suppose; as is, of course, the rounding-off of digits in the computational process.

3.6 Value of information. With decision undelayed, costless and unconstrained, and inquiry undelayed, the problem of the chooser of a two-

link chain $\underline{P}$ is simplified. Denote by $\{\alpha'\}$ the set of <u>all</u> stochastic transformations from Y to A (any such transformation is feasible) and let $\{\eta,\gamma\}$ be the set of feasible pairs of inquiry transformations η and inquiry cost functions γ. Then the constraint (3.3.2) is relaxed into

$$(3.6.1) \qquad \Omega \equiv (\eta,\alpha,\gamma) \in \{\alpha'\} \times \{(\gamma,\eta)\}' ,$$

since $\delta = 0$. Further, equation (3.3.3) is unaffected, but in (3.3.4) the term involving δ vanishes. Therefore, (3.3.5) can be rewritten as

$$(3.6.2) \qquad U = U_{\pi\beta}(\eta,\alpha,\gamma) = B_{\pi\beta}(\eta,\alpha) - C_{\pi}(\gamma) ,$$

where

$$(3.6.3) \qquad B_{\pi\beta}(\eta,\alpha) = \sum_{zya} \beta(a,z)\, \pi_z \eta_{zy} \alpha_{ya}$$

$$(3.6.4) \qquad C_{\pi}(\gamma) = \sum_{z} \pi_z \gamma_z .$$

Define the "<u>information value</u>" of η:

$$(3.6.5) \qquad V_{\pi\beta}(\eta) \equiv \max_{\alpha \in \{\alpha\}} B_{\pi\beta}(\eta,\alpha) \equiv B_{\pi\beta}(\eta,\alpha^*), \quad \text{say;}$$

then, to maximize expected utility U with respect to η, α, γ over their feasible set, given π, β, is equivalent to

$$(3.6.6) \qquad \max_{\eta} V_{\pi\beta}(\eta) - \min_{\gamma} C_{\pi}(\gamma) ,$$

subject to the cost constraint

$$(3.6.7) \qquad (\gamma,\eta) \in \{(\gamma,\eta)\}' .$$

With the meta-decider's problem reduced to (3.6.6), (3.6.7) it is useful to consider the expected cost $C_\pi(\gamma)$ as fixed and to compare various information matrices η, η', ... according to their values $V_{\pi\zeta}(\eta)$, $V_{\pi\zeta}(\eta')$, ...

3.7 Appropriate action, a_y; value of observation, V_y. The optimal decision rule α^* defined in (3.5.5) depends only on η and π, ζ:

$$\alpha^* = \alpha^*_{\pi\zeta}(\eta) \; ; \tag{3.7.1}$$

now, for each η, given π and ζ, there will exist a deterministic optimal decision rule; for it is well-known that, in a one-person game, there exists a pure optimal strategy. Thus, no generality is lost if we define $\{\alpha\}$ as the set of all mappings from $\underline{Y}$ to $\underline{A}$. The assumption of costless and non-restricted decisions excludes the case when the hired (and presumably cheap) decision-making man or machine uses a non-optimal deterministic rule; and also the case when he (it) makes "random errors," unless they happen to constitute an optimal random strategy.

With $\{\alpha\}$ reduced to the set of all pure strategies, i.e., all functions α from $\underline{Y}$ to $\underline{A}$ we can write $a = \alpha(y)$ so that [similar to (1.1.1)]

$$\alpha_{ya} = \begin{matrix} 1 \\ 0 \end{matrix} \quad \text{if} \quad a \begin{matrix} = \\ \neq \end{matrix} \alpha(y) \; ,$$

and denote the action that is "appropriate" (i.e., optimal) in response to $\underline{y}$ by

$$a_y = \alpha^*(y) \; ;$$

that is, for a given y,

$$(3.7.2) \qquad \max_{a \in A} \sum_z \beta(a,z)\pi_z\eta_{zy} \equiv \sum_z \beta(a_y,z)\pi_z\eta_{zy} \equiv V_y ,$$

say. V_y may be called "value of the observation y." It follows by (3.6.3) and (3.6.5) that the value V of an inquiry is the sum of the V_y; for

$$(3.7.3) \qquad V = \max_{\alpha} \sum_z \sum_y \beta(\alpha(y),z)\pi_z\eta_{zy} ,$$

$$(3.7.4) \qquad V = \sum_y \max_a \sum_z \beta(a,z)\pi_z\eta_{zy} ,$$

$$(3.7.5) \qquad V = \sum_y V_y .$$

We shall write $Z = (1,\ldots,m)$, $Y = (1,\ldots,n)$; hence η is of order $m \times n$.

3.8 Labelling of observations. It is clear from (3.6.3), (3.6.5) that $V(\eta)$ is invariant under interchange of columns in η. Therefore, if η is of order $m \times n$ and P a permutation matrix of order n, we shall agree that

$$(3.8.1) \qquad \eta \text{ and } \eta P \text{ are equivalent.}$$

Thus if (with $m = n = 2$), $z = 1$ means "stock will rise" and $z = 2$ means "stock will not rise," then the datum "my broker says stock will rise" can be labelled, indifferently, as $y = 1$ or as $y = 2$. There is no loss of generality in choosing any one particular labelling.

Also, no generality is lost if we agree to eliminate any column of η that consists of 0's only, and thus designates (with Y finite, as

we recall!) an observation that never occurs.

It is seen from (3.7.2) that two observations $y = j, k,$ whose conditional probabilities, given any event z, are pairwise equal, yield the same appropriate action $a_j = a_k$ and the same value $V_j = V_k$. It is convenient therefore, and involves no loss of generality, to redefine every such inquiry by adding any two identical columns, and thus to make every inquiry matrix η to consist of non-identical columns only.

3.9 Null-information is said to be provided by any matrix η whose rows are identical, so that we can write

$$\eta_{zy} = \lambda_y, \quad \text{all} \quad z; \quad \sum_{y=1}^{n} \lambda_y = 1 .$$

Then by (3.7.4)

$$V = \sum_y \lambda_y \max_a \sum_z C(a,z)\pi_z$$

$$(3...1) \qquad V = 1 \cdot \max_a \sum_z C(a,z)\pi_z ,$$

so that V is independent of η. Thus all null-information inquiries have the same value. As their canonical form we can conveniently choose the $(m \times 1)$ matrix with all elements $\eta_{z1} = 1$, $z = 1...,m$. That is, the same unique observation is obtained, with certainty, whatever the event. Then η is the column vector of order m, with all elements = 1: a "sum vector," sometimes denoted by

$$\eta = \underline{1} .$$

3.10 Essential set of inquiry matrices. Let $\{\eta_m\}$ be the set of

all Markov matrices with m rows and with all columns non-zero and not pairwise identical. Summarizing the conventions just made, the essential set H_m of inquiries about m events is defined as the partition $\{\eta_m/e'\}$ into equivalence classes; where η and η' in $\{\eta_m\}$ are equivalent, $\eta = [\eta_{zy}]$ e $\eta' = [\eta'_{zy'}]$ if $\eta' = \eta P$ for some permutation matrix P or if every η_{zy}, $\eta_{zy'}$ is independent of z.

3.11 Perfect information will be said to be provided by a matrix η of order $m \times m$ such that the correspondence between Z and Y is one-to-one. That is, one element in each row of η is $= 1$ (and hence the other elements in the row are $= 0$) and η is noiseless, as in (1.1.1)); and, moreover, in each column one element = 1 and all other elements are = 0. Thus η is a permutation matrix, $\eta = Q$, say. Its transpose Q^T is clearly a permutation matrix, too, $Q^T = P$, say; and it is well known that

$$I = QP,$$

where I is the identity matrix. Then by (3.8.1), I and Q will be considered equivalent: without loss of generality, perfect information will be represented by the identity matrix I as its canonical form.

3.12 Informativeness and optimality of inquiry. In Section 4, a strong partially ordering relation called "more informative than" will be introduced on the essential set H_m of information matrices. This relation is of general significance as it is independent of π and C and is in this sense common to all users (meta-deciders). Some applications to delayed processings will be made in Section 5, still focussing on values V only, by considering expected costs C as given. In

Section 6, $\underline{C}$ will be permitted to vary, to analyze optimality conditions in greater generality.

3.13 Useless inquiries. It will be seen in Section 4.3 that, for any π, β, the value of η cannot be smaller than the value common to all null-information inquiries, given in (3.9.1). An inquiry will be called useless with respect to π, β if its value, as defined in (3.6.5) is equal to the value of a null-inquiry. Thus all null-inquiries are useless. But (as will be shown on an example in Section 6.4), the converse is, in general, not true.

3.14 The information value $V_{\pi\beta}(\eta)$ is a convex function of η. For, by (3.6.3), the benefit $B_{\pi\beta}(\eta,\alpha)$ is linear in its elements η_{zy} of η. Hence, for π, β given, all benefit functions constitute a family of (weakly) convex functions of η. It follows by (3.6.5) and a well-known theorem (see e.g., Karlin [1959], Appendix B.4), that the information value

$$V_{\pi\beta}(\eta) = \max_{\alpha} B_{\pi\beta}(\eta,\alpha)$$

is a convex function of η; it is represented by the upper envelope of a family of hyperplanes.*/ The same is true of V_y in (3.7.2).

3.15 The case of smooth benefit functions. Suppose the set $\underline{A}$ of actions is non-countable, and the benefit function $\gamma(a,z)$ is twice differentiable with respect to $\underline{a}$. Then the observation value V_y and the information value V are continuously differentiable in the elements η_{zy} of η. Moreover it can be conjectured [by extending the reasoning that follows equation (6.5.10)] that in that case all useless inquiries are null-inquiries if $\underline{A}$ is unbounded and there are only two benefit-relevant events.

*/Acknowledgments to a suggestion of M. Pham-Huu-Tri.

4. COMPARATIVE INFORMATIVENESS

4.1 Definition. We say, following Blackwell[1] that η is more[2] informative than η', and write $\eta > \eta'$, if and only if

$$V_{\pi\beta}(\eta) \geq V_{\pi\beta}(\eta') \quad \text{for all } \pi, \beta ,$$

where π, β are defined on fixed sets Z and $A \times Z$, respectively. By fixing these sets rich enough, we can apply the definition of "more informative than" to an arbitrarily large set of meta-deciders concerned with the choice among inquiry matrices, provided the expected cost of information is kept constant.

Clearly ">" is a transitive and reflexive relation, and thus induces an ordering on the set of information matrices. It is a partial ordering on this set: for it is easy to construct cases when, depending on π, β, the information matrix η has a larger or a smaller value than η'. Clearly the relation ">" induces also a partial ordering on the essential set $\{\eta_m/e\}$, defined in Section 3.10. In particular, when $\eta \, e \, \eta'$ then obviously both $\eta > \eta'$ and $\eta' < \eta$. We shall show in Section 4.7 that the converse is also true, so that the partial ordering on the essential set of information matrices by the relation ">" is a strong one.

1/ Several papers by Blackwell and also some earlier work by Bohnenblust, Shapley and Sherman are summarized, as far as "informativeness" is concerned, in Chapter 12 of Blackwell and Girshick [1954]. See also Marschak and Miyasawa [1968].

2/ The "more" (rather than "not less") and the sign ">" (rather than "≥") should not confuse. Blackwell's notation has the advantage of reserving the sign "=" (usually equivalent to "≥ and ≤") for the case of identity. The same would be achieved by symbols "≳" and "∼" used in the economics of preference.

4.2 Garbling. Consider an information matrix $\eta = [\eta_{zy}]$ and suppose that, whenever the observation y $(= 1,\ldots,n)$ is made, the decision-maker does not learn it; instead, a random device is used such that, given the observation y, he will receive, with probability $g_{yy'}$, a signal $y' = 1,\ldots,n'$. Clearly $g_{yy'} \geq 0$, $\sum_{y'} g_{yy'} = 1$. The random device is thus characterized by a Markov matrix $G = [g_{yy'}]$, of order $n \times n'$. It follows that, given the event $z = 1,\ldots,m$, the decision-maker receives signal y' with probability

$$\sum_{y} \eta_{zy} g_{yy'} = \eta'_{zy'} , \tag{4.2.1}$$

say, where $\eta'_{zy'} \geq 0$, $\sum_{y'} \eta'_{zy'} = 1$. In effect, he has used an information matrix $\eta' = [\eta'_{zy'}]$ of order $m \times n'$ such that

$$\eta' = \eta G . \tag{4.2.2}$$

It seems to agree with common usage, to say that η' is obtained from η by garbling. And it is intuitively clear that a garbled information matrix cannot exceed in value the original one: for the decision-maker receiving a "garbled" signal will, at best, choose an action appropriate to that signal, not to the original observation. Formally, we have

Theorem: If η, η', G are Markov matrices with $\eta' = \eta G$, then $\eta' > \eta$.

Proof: By (3.7.2), (3.7.5), (4.2.1)

$$V(\eta') = \sum_{y'z} \sum \ldots (a_{y'}, z)\pi_z \eta_{zy'} = \sum_{y'z} \sum \ldots (a_{y'}, z)\pi_z \sum_{y} \eta_{zy} g_{yy'} =$$

$$= \sum_{y'} g_{yy'} \sum_{y} \sum_{z} \mathcal{J}(a_{y'}, z)\pi_z \eta_{zy} =$$

$$= 1 \cdot \sum_{y} \sum_{z} \mathcal{J}(a_{y'}, z)\pi_z \eta_{zy}$$

$$\leq \sum_{y} \max_{a} \sum_{z} \mathcal{J}(a, z)\pi_z \eta_{zy} = V(\eta) ,$$

by (3.7.4).

4.3 Maximal and minimal information matrices. Theorem:

$$(4.3.1) \qquad \begin{matrix} I_m > \eta \\ \\ \eta > \underline{1}_m , \end{matrix}$$

where η has m rows and I_m and $\underline{1}_m$ (identity matrix and sum vector of order m) correspond to perfect and to null-information (Sections 3.11, 3.9). Proof: Verify that

$$\eta = I_m \eta, \quad \underline{1}_m = \eta \underline{1}_n ,$$

for any η of order $m \times n$; then, noting that η and $\underline{1}_n$ are Markov matrices, apply the Theorem of Sec. 4.2 on "garbling."

Thus perfect information and null information matrices constitute, respectively, the maximal and minimal elements of the lattice in which the essential set of information matrices is partially ordered by the relation "more informative than."

4.4 Comparative coarseness. Suppose the garbling matrix G in

(4.2.2) is noiseless, i.e., analogous to (1.1.1),

(4.4.1) $$\varepsilon_{yy'} = \begin{matrix} 1 \\ 0 \end{matrix} \quad \text{if} \quad y' \begin{matrix} = \\ \neq \end{matrix} g(y),$$

for all y, y'. That is, G is reduced to a many-to-one mapping, g, from $Y = (1,\ldots,n)$ to $Y' = (1,\ldots,n')$; and clearly $n' \leq n$. Then it seems to agree with common usage to say that Y' is coarser than Y (or, equivalently, Y is finer than Y'). For example, two elements y_1 and y_2 may be real numbers (or vectors), identical except for the last digit (or the last component), and this digit (or component) is omitted in the element $y'_1 = g(y_1) = g(y_2)$ of Y'. "Some details are suppressed"; or more generally (to include the limiting case $G = I_n$, $n' = n$), "no details are added." Applying (4.4.1) to (4.2.1),

$$\eta'_{zy'} = \sum_{y \in S_{y'}} \eta_{zy} \quad , \text{ where } S_{y'} = \{ y : g(y) = y' \} :$$

an intuitively obvious result. It follows from the Theorem of Section 4.2 that

(4.4.2) *if η is coarser than η' then $\eta \geq \eta'$.*

This confirms the intuitive assertion that adding detail (at no cost!) cannot do damage, since the detail can be ignored.

4.5 *Blackwell's Theorem.* We give this name to the proposition that

$$\eta \geq \eta' \text{ if and only if } \eta' = \eta G \text{ for some Markov matrix } G .$$

The sufficiency part was proved in Section 4.2. For proof of necessity, see Blackwell [1954] or Marschak and Miyasava [1968].

4.6 The case of noiseless information.

Theorem: If η and η' are noiseless then $\eta > \eta'$ if and only if η' is coarser than η.

Proof: Sufficiency follows from (4.4.2). Necessity follows from Blackwell's theorem, noting that if $\eta' = \eta G$ and η, η' are noiseless then by (4.2.1) every entry in G is either 1 or 0, i.e., G is noiseless. (For a possibly more instructive, direct proof see Marschak and Radner [in press].)

4.7 Strong ordering by informativeness. It can be shown that

(4.7.1) for any two non-null information matrices η, η',

$$V_{\pi\xi}(\eta) = V_{\pi\xi}(\eta') \quad \text{for all } \pi, \xi$$

if and only if η and η' are identical up to a permutation of columns.

The sufficiency part of this proposition is obvious (see also Section 3.8). The necessity part can be restated using the ordering relation ">" and the equivalence relation $\underline{e}$ of Section 3.10, thus:

(4.7.2) If $\eta > \eta'$ and $\eta' > \eta$ then $\eta \underline{e} \eta'$.

It would follow that (as stated at the end of Section 4.1) the partial ordering of the essential set of information matrices by the relation "more informative than" is a strong one.

Outline of proof. The hypothesis of (4.7.2) implies by Blackwell's theorem (Section 4.5, that there exist two Markov matrices G, G' such

that

(4.7.3) $$\eta' = \eta G, \quad \eta = \eta' G' ,$$

and hence

(4.7.4) $$\eta = \eta G G' .$$

We can use two lemmas (proofs omitted). First, to show that $GG' = I$ unless η is null, we use

Lemma 1: If A and B are two Markov matrices and $A = AB$ then B is an identity matrix or A consists of identical rows.

The proof of the theorem is completed by using

Lemma 2: If the product of two Markov matrices is the identity matrix, then they are permutation matrices.

Note: The theorem of this Section is obvious for the case of noiseless information matrices, in view of Section 4.6: for if g maps Y onto Y', and g' maps Y' onto Y, then g and g' must be one-to-one mappings. -- For the general case, I would have liked but have not succeeded to provide a direct proof, not involving Blackwell's theorem and in a sense more instructive: to show that the equality in (4.7.1) cannot be maintained under some well-chosen variations of π, β, except when $\eta \sim \eta'$.

5. INFORMATIVENESS OF SYSTEMS OVER TIME

5.1 Environment, action, and observation as time-sequences. One or both of the arguments $\underline{a}$, $\underline{z}$ of the benefit function β can be interpreted as time-sequences, as in (1.6.1), assuming additive costs as in Section 1.7. With $\underline{z}$ a time-sequence, it will be convenient (changing our terminology somewhat) to call $\underline{z}$ the environment and to reserve the term "successive events" to the components of the sequence $\underline{z} = \{z_t\}$, $t = t_1,\ldots,t_T$; to give unit-length to each of the intervals (t_i, t_{i+1}), $i = 1,\ldots,T$; and sometimes to make $t_1 = 1$, so that $t = 1,\ldots,T$. Each component a_t of $\underline{a}$ will be called successive action. If the benefit can be represented as a sum of discounted "successive benefits"

$$(5.1.1) \qquad \beta(a,z) = \sum_{t=1}^{T} d^t \beta^*(a_t, z_t),$$

say (as would be implied by the assumption (1.6.2) combined with (1.7.3)), then it is important to agree that a_t and z_t need not "physically" occur simultaneously: e.g., a_t may be "sell stock short to-day" and z_t may be "stock price a month from to-day."

A successive action a_t is taken, using the decision rule α_t, in response to $\bar{y}_t$ (note the bar!) where $\bar{y}_t$ is, the remembered past history of successive observations,

$$(5.1.2) \qquad \bar{y}_t = (y_{t-\mu},\ldots,y_{t-1},y_t);$$

the time-length μ measures the length of memory. Again, the subscript t in $\bar{y}_t$ means only that the action taken at time t is based on $\bar{y}_t$; it does not necessarily mean that y_t , the last component of $\bar{y}_t$, was "physically" observed at time t .

In this interpretation, π becomes a distribution on the set $\underline{Z}$ of sequences $\underline{z}$. The information matrix η transforms (stochastically, in general) the environment $\underline{z}$ into a sequence of remembered histories,

$$(5.1.3) \qquad y = (\bar{y}_{t-\mu},\ldots,\bar{y}_T) \in Y ;$$

that is, η_{zy} is the probability of the sequence $\underline{y}$ of remembered histories, given a particular environment (i.e., a particular sequence of successive events), $z = (z_1,\ldots,z_T)$. A strategy α is a sequence of functions $\alpha_1,\ldots,\alpha_T$, where $a_t = \alpha_t(\bar{y}_t)$, thus α is a function from $\underline{Y}$ to the set $\underline{A}$ of action-sequences. (As stated in Section 3.7, the α_t , and thus α , need not be stochastic). With these generalizing interpretations, the results of Section 4 apply.

5.2 Effect of memory length on informativeness. Let $\mu' < \mu$; let inquiry η' yield remembered history

$$(5.2.1) \qquad \bar{y}'_t = (y_{t-\mu'},\ldots,y_t)$$

whenever inquiry η yields remembered history

$$(5.2.2) \qquad \bar{y}_t = (y_{t-\mu},\ldots,y_{t-\mu'},\ldots,y_t);$$

clearly η' is coarser than η. Hence by (4.4.2) η is more informative than η'.

5.3 Delayed vs. prompt perfect information. Prompt perfect and delayed perfect information are defined, respectively by

$$y_t = z_t \ , \ t = 1,\ldots,T$$

$$y'_t = z_{t-\theta}, \ t = \theta + 1,\ldots,T \ ;$$

θ is the delay, an integer with $0 < \theta < T$. Now, there is a one-to-one correspondence between the set $\underline{Z}$ of environments $\underline{z}$ (sequences of successive events) on the one hand, and, on the other, the set, $\overline{Z}$ (say) of sequences $\overline{z} = (\overline{z}_1,\ldots,\overline{z}_T)$ of past histories, $\overline{z}_t = (z_1,\ldots,z_t)$, of successive events: for $\overline{z}_{t+1} = (\overline{z}_t, z_{t+1})$. Replace $\underline{Z}$ by $\overline{Z}$ and redefine β and π accordingly. Then prompt perfect inquiry, η, say, is represented by the identity matrix I; but delayed perfect inquiry is not. Hence $\eta > \eta'$, by (4.3.1). A delay cannot improve perfect information. But if prompt information is not perfect, its value can be exceeded by that of delayed (perfect or imperfect) information. Thus, detailed survey data, even when 2 years old, may be more valuable (because less "coarse": see Section 4.4) than those of a less detailed survey made at the time the action is taken.

5.4 Perfect information with long vs. short delay when the environment is Markovian. Given the distribution π on the set of environments (sequences of successive events) we can derive the conditional probability of the event z_t given the preceding past history, *)

$$\bar{p}_t \equiv p(z_t|\bar{z}_{t-1}) ,$$

and also the conditional probability of z_t given z_{t-1} ,

$$p_t \equiv p(z_t|z_{t-1}).$$

The environment $\underline{z}$ is said to be Markovian if

(5.4.1) $\quad \bar{p}_t = p_t$.

Theorem. If $\underline{z}$ is Markovian then a perfect inquiry with shorter delay is more informative than a perfect inquiry with longer delay.

Outline of Proof. We omit the proof of the following

Lemma: If z is Markovian and $t_1 < t_2 < t_3$ then $p(z_{t_3}|z_{t_2},z_{t_1}) = p(z_{t_3}|z_{t_2})$.

Now let two perfect inquiries, η_θ and $\eta_{\theta'}$, be characterized, respectively, by

(5.4.2) $\quad y_t = z_{t-\theta}$, $y'_t = z_{t-\theta'}$,

where $\theta < \theta'$. If $\underline{z}$ is Markovian then by the Lemma,

(5.4.3) $\quad p(z_t|y_t,y'_t) = p(z_t|y_t)$,

or temporarily omitting the subscript t for brevity,

*) We use the same functional symbol p for various conditional and joint probabilities. $p(\cdot|\cdot)$, $p(\cdot,\cdot)$; no ambiguity arises if one pays attention to the arguments within the parentheses. END OF FOOTNOTE.

$p(z|y,y') = p(z|y)$, that is

$p(z,y,y')/p(y,y') = p(z,y)/p(y)$; hence

$p(z,y,y')/p(z,y) = p(y,y')/p(y)$,

$p(z,y,y')/p(y|z)\cdot p(z) = p(y'|y)$,

$p(z,y,y')/p(z) = p(y|z)\cdot p(y'|y)$,

$p(y,y'|z) = p(y|z)\cdot p(y'|y)$;

summing over y and restoring the subscript t,

$$p(y'_t|z_t) = \sum_{y_t} p(y_t|z_t)\cdot p(y'_t|y_t).$$

Thus inquiry $\eta_{\theta'}$ can be obtained from η_θ by garbling, as in (4.2.1). Hence by the Theorem of Section 4.2, η_θ is more informative than $\eta_{\theta'}$.

As in the case $\theta = 0$ discussed in Section 5.3 for all (not necessarily Markovian) environments, the condition $\theta' < \theta$ does not imply greater informativeness of η_θ compared with $\eta_{\theta'}$, if $\eta_{\theta'}, \eta_\theta$ are not perfect inquiries in the sense of (5.4.1); for then, even if z is Markovian, (5.4.3) would not follow. So that, again, a shorter delay can be profitably traded off against greater precision.

Furthermore, shorter delay is not necessarily advantageous if the environment is not Markovian but is, for example, periodic. Restaurant menus do not vary much as between Sundays, and also, in Catholic countries, as between Fridays. And both differ from each other and from the menus of other days of the week. In a Catholic country, before

deciding on a Thursday where to eat next Sunday, it is best to know next Sunday's menu ($\theta = 0$, as in Section 5.3); but the next best is to learn the menu, not of next Friday ($\theta = 2$ days) but of the previous Sunday ($\theta = 7$ days)!

5.5 Obsolescence and impatience. The discount constant d, as used in Sections 1.6, 1.7, reflects a feature of the utility function, sometimes called impatience. It is one reason why delays diminish the value of an inquiry (and, more generally, of information systems: see end of Section 2.3). We see now another reason, which, when it is applicable, may be more powerful: the obsolescence of the inputs to the decision-making.[1)]

5.6 Sequential inquiries and adaptive programming. The concept a_t of a successive action (decision) can be usefully extended to include decisions about the observations to be taken at the next point of time. Thus

$$(5.6.1) \qquad a_t = (a'_t , \eta_{t+1}),$$

where a'_t may be called, successive action in the ordinary sense (it enters the benefit function) and η_{t+1} is "inquiry at time $t+1$." Both are chosen simultaneously, on the basis

---

[1)] Further analysis, using some special classes of environment distributions π and benefit functions β is given in Chapter 7 of Marschak and Radner [in press].

of remembered history, $\bar{y}_t$. Sequential sampling in statistics is a special case, with a'_t including among its values the null-action: do nothing that would directly influence the benefit, and η_{t+1} including among its values the null-inquiry $\underline{1}$; a'_{t+1} is null, (i.e., ordinary action is postponed) and η_{t+1} is non-null (i.e. further observations are taken), till some point τ (say) such $\eta_{\tau+1}$ is null (observations cease) and a'_τ is non-null ("terminal action"). The more general case is "earn while you learn".

Inquiring and deciding over time, including the general, sequential case just discussed is sometimes called adaptive programming. This is sometimes described as a sequence of step-wise revisionsof the probability distribution of the environment, starting with the prior distribution π and replacing it with posterior distributions, given past histories, $p(z|\bar{y}_t)$, $t = 1,...$ This description can lead to misapplications, if the researcher estimates each of these successive distributions by some conventional parameters (means, variances, for example). The parameter actually needed is the optimal action a^*_t (say) itself! Also, a misleading distinction is sometimes made between "stochastic programming" in which the distribution of $\underline{z}$ is known, and "adaptive programming" in which it is gradually learned. But actually, once the knowledge of the prior distribution π is admitted the mathematical processes needed to compute the optimal sequence of actions (including inquiries as in (5.6.1) are equivalent.[1)]

---

1) See Bellman [1961], Marschak [1963], Miyasawa [1968].

6. OPTIMAL INQUIRIES

6.1 Binary information matrices as an example. The "likelihood matrix" $\eta = [\eta_{zy}]$ is called binary if it is of order 2x2, so that $Z = (1,2)$, $Y = (1,2)$ and we can write

$$\eta_{11} = 1 - \eta_{12} = p_1$$
$$\eta_{22} = 1 - \eta_{21} = p_2 \,. \tag{6.1.1}$$

To avoid triviality, we assume the probabilities π_z of the two events to be both positive:

$$0 < \pi_2 = 1 - \pi_1 < 1 \,. \tag{6.1.2}$$

Binary information matrices are widely used in statistics. In testing against a null-hypothesis, the "error probabilities of first and second kind" are defined as η_{11}, η_{22} or their complements. Binary "channel matrices" are much used in the theory of communication. We shall look to both fields for examples when, later in this section, we compute the maximal difference between expected benefit and expected cost, using sampling costs as well as the cost of a channel.

6.2 Informativeness of binary inquiries. (For brevity, we speak of "inquiries" instead of "information matrices" even though we are, in fact, concerned with the stochastic transformation η characterizing the whole information processing

chain: see Section 2.3). Given the matrix

$$(6.2.1)\quad \eta = [\eta_{zy}] = \begin{pmatrix} p_1 & 1-p_1 \\ 1-p_2 & p_2 \end{pmatrix},\ 0 \le p_i \le 1, i = 1,2,$$

there is no loss of generality in permuting the columns so as to make

$$(6.2.2)\quad p_1 + p_2 \ge 1 .$$

(The "error probabilities of two kinds", usually taken to be small, would then be denoted by $1-p_1, 1-p_2$). Define the two likelihood ratios

$$(6.2.3)\quad \lambda_1 = p_1/(1-p_2);\ \lambda_2 = p_2/(1-p_1).$$

Then under the convention (6.2.2),

$$(6.2.4)\quad \lambda_1 \ge 1,\ \lambda_2 \ge 1 ,$$

and, denoting by $p_y^{(k)}$, $\lambda_y^{(k)}$ $(y = 1,2)$, respectively, the likelihoods and likelihood ratios characterizing the matrix $\eta^{(k)}$, we have the following

Theorem. (1) If $\lambda_y^{(1)} \ge \lambda_y^{(2)}$ $(y = 1,2)$, then $\eta^{(1)} \succ \eta^{(2)}$, and conversely.

(2) If $p_y^{(1)} \ge p_y^{(2)}$ $(y = 1,2)$, then $\eta^{(1)} \succ \eta^{(2)}$, but the converse is not true.

(1) and the first part of (2) follow from Blackwell's theorem (Section 4.5 above). For the second part of (2) let

$$p_1^{(1)} - p_1^{(2)} > p_1^{(1)} p_2^{(2)} - p_1^{(2)} p_2^{(1)} > p_2^{(2)} - p_2^{(1)} > 0 .$$

Then $p_1^{(1)} > p_1^{(2)}$ but $p_2^{(1)} < p_2^{(2)}$; yet $\lambda_1^{(1)} > \lambda_1^{(2)}, \lambda_2^{(1)} > \lambda_2^{(2)}$ so that by (1) $\eta^{(1)} > \eta^{(2)}$.

then condition 6.2.5 is satisfied. Thus, regardless of penalties for errors of first and second kind (i.e., regardless of the benefit matrix β : see Section 6.4) it may pay to decrease the error probability of only one kind while increasing that of the other. (See Figure 1).

It is clear that all null-information matrices (Section 3.9) satisfy

(6.2.5) $p_1+p_2 = 1$; $\lambda_1 = \lambda_2 = 1$, (main diagonal in Fig. 1)

while perfect information is characterized by

(6.2.6) $p_1=p_2 = 1$; λ_1,λ_2 infinite (point (1,1) in Fig. 1)

6.3 Symmetric binary information matrices. This is a special case of (6.2.1), with

$$p_1 = p_2 = p ,$$

say. The convention (6.2.2) becomes

$$p \geq \tfrac{1}{2} ,$$

and it follows from the theorem of the preceding section that the information value is non-decreasing in p: an intuitively obvious result. On Fig. 1, the symmetric matrices are represented by the line (not drawn) connecting $(\frac{1}{2},\frac{1}{2})$ and (1,1).

6.4 Benefit matrix and information value: the case of two actions. As stated in Section 1.8, no two rows and no two columns of the matrix $\beta = [\beta_{az}]$ are identical; and any action represented by a dominated row is eliminated. If after such elimination there remain two rows, i.e. A=(1,2) there is no loss of generality in writing

$$(6.4.1)\quad \beta=[\beta_{az}]=\begin{pmatrix} b_1 & b_2-r_2 \\ b_1-r_1 & b_2 \end{pmatrix},\ r_z>0, z=1,2\ ;$$

the r_z are often called "regrets" (about not having used the action $a=z$, optimal under certainty). This benefit matrix is, in effect, used in statistics when the two actions are: "reject the hypothesis" and "accept it;" the r_z are then penalties for committing an error of first or second kind.

For brevity, write

$$(6.4.2)\quad q_i = 1 - p_i\ ,\ i = 1,2\ .$$

With both η and β of order 2x2 the value of information is, by (3.7.2), (3.7.5)

$$V(\eta) = V_1(\eta)+V_2(\eta),\ \text{where}$$

$$V_1(\eta) = \max\ (\beta_{11}\pi_1 p_1+\beta_{12}\pi_2 q_2,\ \beta_{21}\pi_1 p_1+\beta_{22}\pi_2 q_2)$$

$$V_2(\eta) = \max\ (\beta_{11}\pi_1 q_1+\beta_{12}\pi_2 p_2,\ \beta_{21}\pi_1 q_1+\beta_{22}\pi_2 p_2)\ ;$$

then by (4.3.1), (6.2.6), (6.2.5), (6.4.1), the value of perfect and of null-information are, respectively

$$v^{max} = \pi_1\beta_{11}+\pi_2\beta_{22} = \pi_1 b_1+\pi_2 b_2$$

$$v^{min} = \max(\pi_1\beta_{11}+\pi_2\beta_{12},\ \pi_1\beta_{21}+\pi_2\beta_{22}) = v^{max} - \min(\pi_1 r_1, \pi_2 r_2).$$

(Note: If we considered inquiry costs fixed, the comparison between expected utilities (not: benefits) of inquiries would not be affected by putting $b_1=b_2=v^{max} = 0$. This is usually done in statistics).

The "weighted regrets" $\pi_z r_z, z=1,2,$ are by (6.1.2), (6.4.1) always positive, and we can label the events $\underline{z}$ so that, without loss generality

$$\pi_2 r_2 \geq \pi_1 r_1 > 0 .$$

Then

$$v^{min} = v^{max} - \pi_1 r_1$$

$$V(\eta) = v^{max} - \min(\pi_1 r_1, \pi_1 r_1 q_1+\pi_2 r_2 q_2),$$

by (6.2.2).(6.4.2).

Hence, remembering (6.4.2),

$$(6.4.3)\quad V(\eta) = \begin{cases} v^{min} & \text{if } \pi_1 r_1 p_1+\pi_2 r_2 p_2 \leq \pi_2 r_2 \\ v^{max}-L(\eta) & \text{otherwise;} \end{cases}$$

where $L(\eta)$, the loss due to imperfection of information, is

$$(6.4.4)\quad L(\eta) = \pi_1 r_1(1-p_1)+\pi_2 r_2(1-p_2).$$

Thus all inquiries such that

$$\pi_1 r_1 p_1 + \pi_2 r_2 p_2 \leq \pi_2 r_2$$

have the same value as null-information. They constitute a "useless" indifference set. In the (p_1,p_2) -plane, all other indifference sets are straight lines parallel to the line

$$(6.4.5) \qquad \pi_1 r_1 p_1 + \pi_2 r_2 p_2 = \pi_2 r_2 \ ,$$

which bounds the triangle representing the useless set. See Figure 2. The information value as a function of (p_1,p_2) over the region (6.2.2) is, then, represented by a horizontal plane and an upward sloping plane, intersecting along the line (6.4.5).

If η is symmetrical, $p_1=p_2=p\geq \frac{1}{2}$, the above results become:

$$(6.4.5) \quad V(\eta) = \begin{cases} V^{\min} & \text{if } p \leq \pi_2 r_2/(\pi_1 r_1+\pi_2 r_2) \\ V^{\max}-(\pi_1 r_1+\pi_2 r_2)(1-p) & \text{otherwise.} \end{cases}$$

Thus, the information value of symmetric binary information (in the case of two actions), if plotted against the probability $p(\geq \frac{1}{2})$, consists of a "useless" horizontal segment till p reaches a certain bound; and is a positively sloped straight line for larger p. See Figure 3a.

6.5 The case of more than two actions. If the number of actions exceeds two, the value of a binary inquiry need not be (piece-wise) linear in the probabilities p_1, p_2; and the indifference curves, including the one bounding the "useless" region need not be linear*/. In fact, the indifference curves can become strictly concave (quite unlike those of consumer theory). This can be shown by inserting appropriate numerical

---

*/ The non-linearity of indifference curves in the considered case of more than 2 actions contradicts a statement of L.J. Savage [1962] who obtained parallel straight lines of equal information values in the plane (p_1,p_2), presumably for any number of actions. Consider the following two objects: 1) an inquiry $\eta = (p_1,p_2)$ (i.e., a binary inquiry with $\eta_{11} = p_1$, $\eta_{22} = p_2$), and 2) a gamble, which we shall denote by $g = (\eta',\eta'';\alpha)$, and which gives you access to inquiries $\eta' = (p_1',p_2')$ and $\eta'' = (p_1'',p_2'')$, with odds $\alpha:(1-\alpha)$. Savage considers, in effect, η and g as identical objects, provided

$$(*)\quad p_1 = \alpha p_1' + (1-\alpha)p_1'', \quad p_2 = \alpha p_2' + (1-\alpha)p_2'' .$$

It is true that, if η' and η'' have equal values, then g has the same value. For a decider in possession of g will respond to observation y by the actions a_y' and a_y'' with respective probabilities α and $1-\alpha$; where a_y' and a_y'' denote the actions appropriate to y when the inquiry is η' or η'', respectively. Hence, for the possessor of g, observation y has value [in the sense of (3.7.2)]

$$V_y(g) = \alpha V_y(\eta') + (1-\alpha)V_y(\eta''), \quad y = 1,2 ,$$

Therefore $V(g) = V_1(g) + V_2(g) = \alpha[V_1(\eta')+V_2(\eta')] + (1-\alpha)[V_1(\eta'')+V_2(\eta'')]$

$= \alpha V(\eta') + (1-\alpha)V(\eta'')$; so that, if $V(\eta') = V(\eta'') \equiv V$, say, then indeed $V(g) = V$.

If, in addition, condition (*) would imply that y and η are identical objects, then it would indeed imply

$$V(\eta) = V(g) = V = V(\eta') = V(\eta'') ,$$

so that the points representing η, η', η'' would lie on the same indifference line. This line would be straight since by (*), (p_1,p_2) lies on a straight line between (p_1',p_2') and (p_1'',p_2''). And all such lines are parallel since you would, by the same reasoning, be indifferent between gambles such as $(\eta',\eta^*;\alpha)$ and $(\eta'',\eta^*;\alpha)$.

However, I don't think that condition (*) makes the objects η and g identical, or their values equal. The possessor of η will respond to observation y by some appropriate action a_y (say) which has, in general, no relation to the actions a_y' and a_y'' which are appropriate when the inquiry is η' or η'', respectively. There is therefore no necessary equality between $V(\eta)$ and $V(g)$, and hence none between $V(\eta)$ and $V(\eta')$ and $V(\eta'')$.

END OF FOOTNOTE

---

values into the 3 x 2 benefit matrix (3 treatments, each with cancer present or absent), mentioned in Section 3.4. To permit the use of calculus consider, instead, a case in which actions constitute a closed non-countable set, $0 \leq a \leq 1$, and the benefit function $\beta(a,z)$, now written $\beta_z(a)$ for convenience, is twice differentiable with respect to a. As before, $Z = (1,2)$. Let $\beta_1(a)$ increase, and $\beta_2(a)$ decrease, in a. Then $a > a'$ implies $\beta_1(a) > \beta_1(a')$ and $\beta_2(a) < \beta_2(a')$, hence no action is dominated.

The following example will impose some further constraints. A farmer wishes to maximize the amount harvested. He must decide on how to allocate his total acreage (=1) between two crops, the "wet" crop being favored by wet weather (denoted by $z = 1$), and the "dry" crop by dry weather ($z = 2$). The action a is the acreage allotted to the wet crop. Let the harvest from c acres of a given crop when the weather is or is not favorable to it, be, respectively, $f(c)$ and $g(c)$. The

combined harvest of the wet and the dry crop is then

$$\beta_1(a) = f(a) + g(1-a) \quad \text{in wet weather}$$
$$\beta_2(a) = f(1-a) + g(a) \quad \text{in dry weather}.$$

It is natural to assume that β_1 increases, and β_2 decreases, in a, so that no action is dominated, and

$$\beta_1'(a) = f'(a) - g'(1-a) > 0; \; \beta_2'(a) = -f'(1-a) + g'(a) < 0 .$$

Finally, if both crops obey the "law of decreasing marginal returns," $f''(c) < 0$, $g''(c) < 0$, then

$$\beta_1''(a) = f''(a) + g''(1-a) < 0; \; \beta_2''(a) = f''(1-a) + g''(a) < 0 ;$$

write

$$\gamma_z(a) \equiv \pi_z \beta_z(a), \quad z = 1,2 ; \tag{6.5.1}$$

then, since $\pi_z > 0$,

$$\gamma_1'(a) > 0; \; \gamma_2'(a) < 0$$

$$\gamma_z''(a) < 0 , \quad z = 1,2 . \tag{6.5.2}$$

This is, in fact, the only additional constraint we need, to show that the information value $V(\eta)$ is strictly convex (and therefore strictly quasi-convex); that is, as p_1, p_2 vary, the second differential $d^2V > 0$ (and therefore the indifference lines are concave.*/) A sufficient condition

---

*/ A twice-differentiable and increasing function $F(x)$ of a vector x is called strictly concave over its sub-domain X if, in that sub-domain, $d^2F < 0$; clearly such a function is also strictly quasi-concave over X, i.e., $d^2F < 0$ holds, in particular, for all x in X such that $dF = 0$. The contour lines of equal values of a quasi-concave F are convex: see, e.g., Arrow and Enthoven [1961]. Now, replacing "<" by ">", the definitions of F strictly convex and strictly quasi-convex (with contour lines concave) are obtained.

for this is:

$$(6.5.3) \qquad w_{11} > 0,\ w_{22} > 0,\ \begin{vmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{vmatrix} > 0 ,$$

where $(6.5.3')$ $\qquad w_{ij} \equiv \partial^2 V / \partial p_i \partial p_j$.

To evaluate the w_{ij}, note that the expected benefit of action $\underline{a}$ when the observation is $\underline{y}$ is

$$B_y(a) \equiv \sum_{z=1}^{2} \gamma_z(a)\, \eta_{zy}, \quad y = 1,2 ;$$

the value of the observation $\underline{y}$ is

$$V_y = \max B_y(a) = B_y(a_y) = \gamma_1(a_y)\eta_{1,y} + \gamma_2(a_y)\eta_{2,y}, \quad y = 1,2 ;$$

$$(6.5.4) \qquad V_y' \equiv \left.\frac{dB_y(a)}{da}\right]_{a=a_y} = \gamma_1'(a_y)\eta_{1,y} + \gamma_2'(a_y)\eta_{2,y} = 0 ,$$

since by (6.5.2) and with all η_{zy} positive,

$$(6.5.5) \qquad V_y'' \equiv \left.\frac{d^2B_y(a)}{da^2}\right]_{a=a_y} = \gamma_1''(a_y)\eta_{1,y} + \gamma_2''(a_y)\eta_{2,y} < 0 .$$

In terms of p_1, p_2, and emphasizing that V_y depends on a_y,

$$(6.5.6) \qquad V_1 = V_1(a_1) = \gamma_1(a_1)p_1 + \gamma_2(a_1)(1-p_2);\ V_2 = V_2(a_2) = \gamma_1(a_2)(1-p_1) + \gamma_2(a_2)p_2 ;$$

$$(6.5.7) \qquad V_1' = V_1'(a_1) = \gamma_1'(a_1)p_1 + \gamma_2'(a_1)(1-p_2) = 0;$$

$$V_2' = V_2'(a_2) = \gamma_1'(a_2)(1-p_1) + \gamma_2'(a_2)p_2 = 0 .$$

Write $\partial a_j/\partial p_i \equiv a_{ji}$; then differentiating (6.5.6) with respect to p_j using (6.5.7)

$$\partial V_i/\partial p_i = V_i'(a_i)a_{ii} + \gamma_i(a_i) = 0 + \gamma_i(a_i)$$

$$\partial V_j/\partial p_i = V_j'(a_j)a_{ji} - \gamma_i(a_j) = 0 - \gamma_i(a_j) \; ; \quad i \neq j \; ;$$

and since $V = V_1 + V_2$,

$$\partial V/\partial p_1 = \gamma_1(a_1) - \gamma_1(a_2); \; \partial V/\partial p_2 = -\gamma_2(a_1) + \gamma_2(a_2) \; ;$$

then by (6.5.3') and writing $\gamma_i'(a_j) = \gamma_{ij}$,

$$\begin{aligned} w_{11} &= \gamma_{11}a_{11} - \gamma_{12}a_{21} \qquad w_{12} = \gamma_{11}a_{12} - \gamma_{12}a_{22} \\ w_{21} &= -\gamma_{21}a_{11} + \gamma_{22}a_{21} \qquad w_{22} = -\gamma_{21}a_{12} + \gamma_{22}a_{22} \; ; \end{aligned} \tag{6.5.8}$$

(it will be confirmed presently that $w_{12} = w_{21}$). To evaluate the a_{ji}, differentiate with respect to p_i the equations (6.5.7), which are identities in the a_j:

$$\begin{aligned} \partial V_1'/\partial p_1 = 0 &= V_1'' \cdot a_{11} + \gamma_{11} \qquad \partial V_1'/\partial p_2 = 0 = V_1'' a_{12} - \gamma_{21} \\ \partial V_2'/\partial p_1 = 0 &= V_2'' \cdot a_{21} - \gamma_{12} \qquad \partial V_2'/\partial p_2 = 0 = V_2'' a_{22} - \gamma_{22} \; , \end{aligned}$$

solve for the a_{ji}, writing for brevity

$$V_j'' = -1/k_j < 0, \quad j = 1,2 \quad \text{[by (6.5.5)]}; \tag{6.5.9}$$

then

$$a_{jj} = -k_j\gamma_{jj}; \; a_{ji} = k_j\gamma_{ij}, \quad i \neq j \; ;$$

and by (6.5.8), (6.5.9),

$$w_{11} = -\gamma_{11}^2 k_1 - \gamma_{12}^2 k_1 \qquad w_{22} = -\gamma_{21}^2 k_1 - \gamma_{22}^2 k_2 > 0$$

$$w_{12} = \gamma_{11}\gamma_{21} k_1 + \gamma_{12}\gamma_{22} k_2 = w_{21}$$

$$w_{11} w_{22} - w_{12}^2 = k_1 k_2 (\gamma_{11}\gamma_{22} - \gamma_{12}\gamma_{21})^2 > 0 ,$$

thus establishing condition (6.5.3), sufficient for $V_1 + V_2$ to be strictly convex in p_1, p_2.

$\underline{V}$ must not be less than the value of null-information which is

$$V^{\min} = \max_a [\gamma_1(a) + \gamma_2(a)] = \gamma_1(a^*) + \gamma_2(a^*) ,$$

where

(6.5.10)
$$\gamma_1'(a^*) + \gamma_2'(a^*) = 0 ,$$

$$\gamma_1''(a^*) + \gamma_2''(a^*) < 0 \quad \text{by (6.5.2)} .$$

In an inquiry is useless, $a_1 = a_2 = a^*$. Then by (6.5.7), (6.5.10), $p_1 + p_2 = 1$. Thus the set of useless inquiries coincides with that of null-inquiries, represented by the main diagonal of the unit-square. (See conjecture in Section 3.15.)

For a simple example, let the prior weather probabilities be $\pi_1 = \pi_2 = \frac{1}{2}$ and let the production functions g, f (for the crop not favored or favored, respectively, by weather) be

(6.5.11)
$$g(c) = (3c - c^2)/8; \quad f(c) = 3g(c); \quad c = a \text{ or } 1-a .$$

Then
$$\beta_1(a) = 2\gamma_1(a) = -\tfrac{1}{2}a^2 + a + \tfrac{1}{4}; \quad \beta_2(a) = 2\gamma_2(a) = -\tfrac{1}{2}a^2 + \tfrac{3}{4} .$$

Apply this to (6.5.6), and solve (6.5.7) for the $a_y, y = 1,2$ (optimal acreages of wet crop), writing $q_i \equiv 1-p_i$. Then

$$a_1 = p_1/(p_1 + q_2); \quad a_2 = q_1/(q_1 + p_2);$$

$$V_1 + V_2 = p_1^2/4(p_1 + q_2) + q_1^2/4(q_1 + p_2) + \tfrac{1}{2}. \tag{6.5.12}$$

It is easily seen that

$$V_1 + V_2 = V^{min} = 5/8$$

if and only if $p_1 + p_2 = 1$. Thus all useless inquiries are null-inquiries, represented by the diagonal $p_1 + p_2 = 1$. All indifference lines in the space above the diagonal are strictly concave since $V_1 + V_2$ in (6.5.12) is strictly convex in p_1, p_2 when $p_1 + p_2 > 1$.

In the symmetric case, i.e., with $p_1 = p_2 \equiv p \equiv 1-q \geq \tfrac{1}{2}$, this example yields

$$a_1 = p, \quad a_2 = q$$

$$V = V_1 + V_2 = (p^2 + q^2)/4 + \tfrac{1}{2},$$

so that, plotted against p, the information value V is represented by a strictly convex, rising curve (a parabola with a minimum at $p = \tfrac{1}{2}$). See Figure 3b.

6.6. Cost conditions. So far, we have explored, at least for the case of binary information matrices, the behavior of the information value function $V(\eta)$ which associates each η with the maximum expected benefit. If utility can be represented as the difference between benefit and information cost, an optimal matrix η maximizes the difference between $V(\eta)$ and the expected information cost, subject to a constraint on feasible pairs (η,γ) of inquiries and cost functions (Section 3.6),

$$(\eta,\gamma) \in [(\eta,\gamma)].$$

A simple assumption is to associate each η with just one cost function $\gamma_z(\eta)$, viz., the one giving the lowest expected cost [as in (3.4.6)], for a given η. In addition we shall make $\gamma_z(\eta)$ independent of z:

$$\gamma_z(\eta) = \gamma(\eta),$$

say. Thus, if η is obtained by a sampling survey of families, the cost $\gamma(\eta)$ will depend on the size of the sample needed to obtain η (i.e., to attain some preassigned error probabilities); but not on the properties of the families -- disregarding, for example, the fact that households of certain types may require second visits.

Using these simplifying assumptions, and still confining ourselves to binary information matrices, we shall give two examples illustrating the possible behavior of the cost functions $\gamma(\eta)$. An important question is: under what conditions does the expected utility, as a function of η,

$$(6.6.1) \qquad U(\eta) = V(\eta) - \gamma(\eta)$$

behave in such a way that the optimal information matrix is an "interior solution". If it does not, the optimal binary information matrix may be the perfect information, $(p_1,p_2) = (1,1)$: a case of "large scale economics," making the competitive market equilibrium non-optimal from the point of view of social welfare. Thus when $V(\eta)$ is quasi-convex, i.e. the indifference curves are concave, the existence of interior solution requires that the lines of equal cost be also concave, with even larger curvature. This requirement would mean, in the case of binary symmetric matrices, with $p(\geq \frac{1}{2})$ replacing η in (6.6.1) in an obvious manner, that,

$$(6.6.2) \quad V'(p) = \gamma'(p) \text{ should imply } V''(p) < \gamma''(p).$$

6.7 Cost linear in channel capacity. The capacity $C = C(p_1,p_2)$ of a channel transmitting one bit per time unit (see below, Section 7.4) is given*) by

$$C = 2^{(p_2H_1 - q_1H_2)/(q_1-p_2)} + 2^{(p_1H_2 - q_2H_1)/(q_1-p_2)},$$

where $$H_i = -(p_i \log_2 p_i + q_i \log_2 q_i), \; i = 1,2 .$$

*) See, e.g. Ash [1965], Theorem 3.3.3 (p.56) and problem 3.7 (p.304). $\underline{C}$ is the conventional symbol for channel capacity. In Section 3.2, $\underline{C}$ was introduced to denote expected information cost, which we shall here assume linear and increasing in channel capacity. Apologies to the reader of this mimeographed paper for this inconsistence in notations. In this section, cost and expected cost are both $= \gamma(p)$.

$\underline{C}$ is quasi-convex in (p_1,p_2). The contour lines of equal capacity are strictly concave for $p_1+p_2 > 1$; all points on the straight line $p_1+p_2 = 1$ have equal capacity $C = 0$; and maximum capacity is $C(1,1) = 1$. See Fig. 4.

Suppose that the indifference lines (contour lines of equal information value) are strictly concave, as in the "farmer's case" of Section 6.5. Suppose further that the 'observations" $y = 1,2$ are messages ("wet", "dry") received through a channel whose inputs are the "true" events (viz. actual future weather), $z = 1,2$. And suppose information cost increases linearly with channel capacity. An optimal information system (consisting in this case of the channel and nothing else) is an "interior" optimum if the optimal pair (p_1,p_2) is such that

$$\underline{\text{not}} : \quad p_1+p_2=1, \text{ or } p_1=1, \text{ or } p_2=1.$$

This requires that the contour lines of equal capacity be "more concave" (i.e. have greater curvature) than the indifference lines.

In the symmetric case, $p_1 = p_2 \equiv p = 1-q$, the channel capacity is

$$C = 1 + p \log_2 p + q \log_2 q \ .$$

If the channel cost $\gamma(p)$ (measured in our farmer's harvest bushels or dollars) is increasing linearly in $\underline{C}$ then the

expected utility is

$$U = V - r\,C - \text{const.} \quad (r > 0)$$

$$U = -\tfrac{1}{2}(1-p) - r[p \log_2 p + (1-p)\log_2(1-p)] - s ,$$

say. It will depend on the constants r,s, whether condition (6.6.2) is fulfilled and thus an interior solution exists.

6.8 Cost of inferring sign of mean of finite population from sign of mean of sample. Suppose n random variables $u_i (i=1,\ldots,n)$ are jointly normal, with

$$E(u_i) = 0, \quad E(u_i u_j) = \begin{cases} 1 & \text{if } j = i \\ 0 & \text{if } j \neq i . \end{cases}$$

Define the events z and the "observations" (usually called "statistics") y by

$$z = \begin{cases} 1 \\ 2 \end{cases} \text{ if } \sum_1^n u_i \begin{cases} > \\ \leq \end{cases} 0 \; ; \quad y = \begin{cases} 1 \\ 2 \end{cases} \text{ if } \sum_1^m \begin{cases} > \\ \geq \end{cases} 0 ,$$

$$1 \leq m \leq n ;$$

thus m is the size of the sample, and n is the size of the population. Then (see Cramèr [1946], p.290) the joint distribution of z and y is given by

$$\Pr(z=1, y=1) = \Pr(z=2, y=2) = 1/4 + (\arc\sin \rho)/2\pi$$

$$\Pr(z=1, y=2) = \Pr(z=2, y=1) = 1/4 - (\arc\sin \rho)/2\pi ,$$

where $\rho = \sqrt{m/n}$ *). Hence η is binary symmetric, with

$$\eta_{11} = \Pr(y=1 \mid z=1) = \tfrac{1}{2} + (\arc\sin\sqrt{m/n})/\pi = \eta_{22} = \rho > \tfrac{1}{2} ;$$

*) See Marschak [1964], equation (35). The example given in that paper has n stocks in a portfolio and a sample of m

of them . It seems more difficult to think of a finite population (finite number of possible future states of humidity) in the case of our "farmer." With population infinite the optimal sample size is convex in $\eta_{11} = p$ (instead of exhibiting an inflexion), so that the example would not add much to that of the preceding Section. END OF FOOTNOTE

$$m = n \sin^2 \pi(p-\tfrac{1}{2})$$

$$dm/dp = \pi n \sin \pi(2p-1)$$

$$d^2m/dp^2 = 2\pi^2 n \cos \pi(2p-1) \gtrless 0 \text{ if } p \lessgtr 3/4 .$$

The sample size $\underline{m}$ is thus an increasing function of $\underline{p}$, convex for small (and hence less informative: Section 6.3) values of $\underline{p}$, and concave for larger ones. So is the cost of information (sampling cost) if we assume it to increase linearly with $\underline{m}$. Therefore a whole range of sufficiently small values of $\underline{p}$ (and therefore of $\underline{m}$) is non-optimal, especially if information $V(p)$ is strictly convex in $\underline{p}$.

7. ECONOMICS OF COMMUNICATION

7.1 The fidelity criterion as benefit. In the preceeding three Sections, the benefit $\beta(a,z)$ depends on the "action" a in A and the "event" (or hypothesis*) z in Z. A probability function π is defined on Z. Event z is transformed into "observation" y by a processing η; and y is transformed into a by a subsequent processing called strategy (these processings are possibly stochastic).

Now let us interpret, instead, z in Z as a "message sent", occurring with probability π_z. Interpret processing η as "communication" (to be specified later as a chain: storing, encoding, transmitting): it transforms message z into y, the latter to be interpreted as some signals received by the decision-maker. An important restriction is this: the set A of actions a is identical with the set Z of messages sent. The strategy α consists then in a rule of "decoding" the received signals y, i.e., in prescribing which element a of Z (or which conditional distribution of a) should be associated with a given y.

The early writings on communication theory -- most importantly the pioneering work of Shannon [1948] - imposed a further restriction, by assuming equal penalty for all communication errors, so that "a miss is as bad as a mile." That is, the benefit function is taken to be simply

$$(7.1.1) \qquad \beta(a,z) = \begin{matrix} 0 \\ -1 \end{matrix} \quad \text{if} \quad a \begin{matrix} = \\ \neq \end{matrix} z ,$$

*See second paragraph of Section 3.1

i.e., $\beta(a,z) = 1$ minus Kronecker delta. Then the expected benefit is, by (3.6.3) and (7.1.1)

$$B_{\pi\beta}(\eta,\alpha) = \sum_z \sum_y \beta(a,z)\ \pi_z\ \eta_{zy}\ \alpha_{ya} =$$

$$= - \sum_{z \neq a} \sum_y \pi_z\ \eta_{zy}\ \alpha_{ya} = - p_e\ ,$$

where p_e denotes the "probability of error". For a given set $\underline{Z}$ (characterized by π), p_e depends on the properties of the communication processing η and the decoding strategy α.

However, the special restriction (7.1.1) was abandoned later, when Shannon [1960] introduced the "fidelity criterion" (and its negative, the "distortion"), a general real-valued function of the message sent and the message decoded. This function is identical with our general benefit function that maps $Z \times A$ into reals; except for the restriction (mentioned above) that replaces $Z \times A$ by $Z \times Z$. A fidelity criterion does, then, assign different penalties (negative benefits) to different errors of communication and decoding. This idea has not yet penetrated the bulk of literature, certainly not the textbooks, on communication theory.*)

*) But see, more recently, Jellinek [1963] and Pham-Huu-Tri [1968]. The coding procedures recommended by Shannon to maximize expected fidelity can be made more efficient in several respects. END OF FOOTNOTE

7.2 Capacity of noiseless channel. We mentioned in Section 3.2 that statistical decision theory neglects delays

in processing. Communication theory does not neglect them. Concepts like the speed of a processing(thruput per time unit), and the maximum of this speed, achievable with a given processing instrument and called its capacity, arise naturally. As a simple case, imagine a noiseless transmission channel. Its inputs/sequences of symbols such as dots and dashes, or numerical digits. Let us call them digits. They are the outputs of the preceding processing link, the encoding, to be discussed in the next Section, 7.3. The digits are transmitted through the channel one by one and received at the other end with no distortion. If the channel is a cable consisting of several wires, several symbols can be transmitted simultaneously. We can therefore diminish delays by increasing the number of wires, which thus measures the channel's capacity: the maximum number of digits that can be transmitted per unit of time.

Channel capacity--already in the noiseless case--is economically significant for two reasons. First, if the inflow of input digits per time unit exceeds the channel capacity, untransmitted, and therefore useless, inputs will pile up indefinitely, with an obvious detriment to the expected benefit. Second, any further increase of capacity, in excess of the inflow of inputs, will diminish the delay between input and output of the channel.

Why delays can diminish expected benefit, is due to "impatience" (preference for early results of actions) as well as to the obsolescence of data--i.e., in our case, of the channel outputs,-on which the choice of action is based. This was discussed in Sections 1. and 5.

While increased channel capacity thus increases expected benefit, it will, in general, also require an increase in cost.

Expected benefit is diminished by delay. But benefit is not necessarily a linear function of delay. Hence (see Appendix I) expected utility (difference between expected benefit and expected cost) is not monotone in expected delay. Therefore, it is not correct to present the economics of communication--even in the simplest case of a noiseless channel--as that of minimizing expected cost for a given expected delay, or expected speed of transmission. Yet, just this seems to be done, in this or similar contexts, in much of the literature, where, essentially, the problem is presented as that of determining an efficient set in the space of expectations of various "criteria." *)

---

*) The clearest formulation of such an efficient set is given by Wolfowitz [1961], in the context of optimal coding for a noisy channel. It seems that the assumption of utility linear in its criteria is implicit in the discussion of optimal design in many fields of engineering. See, e.g. English [1968].
END OF FOOTNOTE

7.3 Minimum expected length of code word, as the "uncertainty at source." If only two possible messages z (=1 or 0, say) can be sent, each can be encoded as a single binary digit, to be transmitted through the channel. However, if a time sequence of T such two-valued messages is to be communicated, less than T digits (and hence less than one digit per message) will be needed on the average if one uses "code words" (binary sequences) with few digits for the more probable and with more digits for the less probable sequence of messages. For example, if one uses this principle and if the odds for z taking its two values are 9:1, then, even if the sequences of messages occur independently ("have no pattern"), it is possible to devise codes which will use, on the average, approximately only .64 or .53 digits per message when $T = 2$ or $T = 3$, respectively. In general, as established by "Shannon's first theorem," the minimum expected length of the code word decreases as T increases, and it converges towards the (never negative) quantity

$$(7.3.1) \qquad -\sum_{z\in Z} \pi_z \log_2 \pi_z \equiv H(\pi), \quad \text{also written as } H(Z) .$$

This limit is valid not only for the case of two-valued messages (as in our example, with $H(\pi) = .47$) but for a set Z of any size m. Since $H(\pi)$ is largest when all the m elements of Z are equiprobable [so that every $\pi_z = 1/m$, and $H(\pi) = m$], the name "amount of uncertainty" (about z) occasionally given to $H(\pi)$ is indeed a suggestive one. Alternatively, one says that $H(\pi)$ units of information are gained if this uncertainty is removed (by learning the actual value of z). Indeed

$H(\pi)$ has been proposed as a "measure" of uncertainty, or of information, because it is additive in the following sense. Let π', π'' characterize two statistically independent sets Z' and Z''; that is, the joint occurrence $z = (z' \text{ and } z'')$ of given messages from the two sets occurs with probability

$$\pi_z = \pi'_{z'} \cdot \pi''_{z''} ;$$

then, by the definition (7.3.1) of the distribution parameter H,

$$H(\pi) = H(\pi') + H(\pi'') . \tag{7.3.2}$$

Similar additivity properties are derived for certain related distribution parameters (such as "uncertainty removed by transmission," of which more later). Since $H(\pi)$ measures the average length of a sequence of binary digits, the measurement unit of "uncertainty" (or its negative, "information") is called, briefly, a bit, following a suggestion of J.W. Tukey.

It is not clear, however, for what economic purpose one should measure uncertainty, or information. Because of the additive property (7.3.2) of the distribution parameter H, specialists in various fields (mathematics, statistics, psychology) expressed enthusiasm: the subtle, intangible concept of information has now become measurable "in a way similar to that as money is used in everyday life" (Rényi [1966]). Indeed a paper currency bill can be measured by the number of dollars it represents, and thus by the amount of some useful commodity at a given price. But it can be also measured (a peso and a hundred peso bill alike) in square inches of its area. If I use it for papering my walls, the latter not the former measurement is appropriate!

Somewhat anticipating the subsequent more detailed discussion, note that a distribution parameter such as $H(\pi)$ cannot alone determine the information value of a system. For $H(\pi)$ depends only on the distribution π, not on the benefit function β. To be sure, the special assumption (7.1.1) of equal penalty for all communication errors does remove variations of the benefit function. This fact may have been the source of misunderstandings about the economic significance of the number of bits gained or lost, regardless of the use the decision-maker can make of them. If a general fidelity criterion (presumably reflecting the decision-maker's needs) is introduced, $H(\pi)$ fails to determine the information value of the system.

What is economically important about $H(\pi)$ is its meaning as the lower limit of the expected length of a code word, given the distribution π. For, the shorter a code word the less is, presumably, the time needed to transmit it, digit by digit; and therefore, for reasons just stated in Section 7.2, the larger the expected benefit.*/

---

*/ Wolfowitz [1961] writes that the function H should

> "for convenience and brevity have a name. However, we shall draw no implicit conclusions from its name, and shall use only such properties of H as we shall explicitly prove. In particular, we shall not erect any philosophical systems on H as a foundation. One reason for this is that we shall not erect any philosophical systems at all, and shall confine ourselves to the proof of mathematical theorems,"

namely, theorems on optimal coding. The present writer, though guided by economic rather than mathematical interest, tends to agree.

On the other hand, note that, to bring the expected length of code words down close to its lower limit, $H(\pi)$, one may have to wait till a very long sequence of messages (T large) is piled up. The resulting delay may offset the acceleration due to the shortening of code words. In addition, there are storage costs.

We can now refer to the "four-link" chain (b) of Section 2.2 . Messages to be sent are stored, encoded, received, and decoded. The benefit (fidelity criterion) depends on the messages to be sent and on the decoded messages; the expected benefit will depend on the probability distribution π characterizing the source (i.e., the messages to be sent) and on the Markov matrices characterizing consecutive processings. Costs and delays arise at each processing link, and their distribution (and hence expectation) depends, too, on π and those Markov matrices.

We have, however, just remarked that the four-link chain is merely a part of the total information system, in which benefit depends on events and actions. Events are transformed, by inquiry, into observations ("data"). These are the messages to be sent, the initial input of the communication system; and its final output, the decoded messages, are transformed into actions by applying strategies. We have thus added two links, one at each end of the communication chain. It remains true that the probability distribution (and hence the expectation) of benefits, costs, and delays depends on the initial distribution π (now attached to events, not to messages received) and on the successive Markov matrices.

We can also regard a communication system as a special case of the

general information system; viz., one in which the processing of events into data and the processing of decoded messages into action are characterized by identity transformations and by zero-costs and zero-delays.

7.4 Noisy channel: transmission rate and capacity. To concentrate on the properties of a channel, it will be convenient to reinterpret our notational symbols again. Let us now designate channel inputs by $\underline{z}$ in $\underline{Z}$, and its outputs by $\underline{y}$ in $\underline{Y}$, analogous to the "events" and "observations" of Sections 3-5. Channel inputs $\underline{z}$, the digits of the encoded message, occur with probabilities π_z. Channel outputs, $\underline{y}$, the digits received at the channel's end, occur, for a given $\underline{z}$, with conditional probabilities $p(y|z) = \eta_{zy}$, elements of the Markov matrix η, called the channel matrix. The channel is noiseless if η is the identity matrix. The joint probability of $\underline{z}$ and $\underline{y}$ and the marginal probability of $\underline{y}$ are, respectively (see footnote to Section 5.4, on notations),

$$p(z,y) = \pi_z \eta_{zy}$$

$$p(y) = \sum_{z \in Z} \pi_z \eta_{zy} .$$

It will be convenient to give a special symbol, δ_{yz}, (an element of the Markov matrix $\delta = [\delta_{yz}]$) to the posterior probability of $\underline{z}$, given $\underline{y}$. Clearly δ depends on π and η:

$$\delta_{yz} = p(z|y) = p(z,y)/p(y) =$$

$$= \pi_z \eta_{zy} / \sum_{u \in Z} \pi_u \eta_{uy} .$$

We may call "uncertainty about $\underline{z}$, retained after digit $\underline{y}$

was received through the channel", the expression

$$H(Z|y) = -\sum_z \delta_{yz} \log \delta_{yz} ,$$

and to call its expectation

$$(7.4.1) \qquad -\sum_y p(y)H(Z|y) = H(Z|Y) ,$$

the "uncertainty retained," in L. Breiman's [1960] suggestive language. It is clear from its definition that $H(Z|Y)$ depends only on the probability distribution π and η, and we want to emphasize this by writing occasionally

$$H(Z|Y) \equiv J(\pi,\eta) .$$

The quantity (never negative)

$$(7.4.2) \qquad H(Z) - H(Z|Y) \equiv I(Z,Y) \equiv I(Y,Z)$$

has been called "uncertainty removed" or "amount of information transmitted." Because of the symmetry with respect to Z, Y, which is easily shown, it has also been called "mutual information."*/ Clearly, it depends on π

---

*/ H. Theil [1967], [1968], uses the difference $H(Z) - H(Z|Y)$ to measure, for example, the discrepancy between the predicted and the actual composition of a balance /sheet, the national income, or some other total. Of course, this measure can be used outside of economics as well; and it is related to information mainly because the same formula has been used in the theory of communication as developed by C. Shannon and others. This explains the difference in content between Theil's studies and those presented here, in spite of the similarity of titles.

END OF FOOTNOTE

and η only, and it will be convenient to write

$$H(\pi) - J(\pi,\eta) = K(\pi,\eta) \quad , \tag{7.4.3}$$

say. Shannon's "generalized first theorem" states that $K(\pi,\eta)$ is the lower limit of the expected number of binary digits, needed to identify (by appropriately decoding the digit sequence received) each digit put through the channel. Thus $K(\pi,\eta)$ is measured by a number of bits, divided by the number of digits put through the channel.

In Section 7.2, the speed of a channel, v digits per time unit (say) was introduced. If we multiply it by $K(\pi,\eta)$ bits per digit, we obtain

$$v(\text{digits/time}) \times K(\pi,\eta)(\text{bits/digit}) = v \cdot K(\pi,\eta)(\text{bits/time}) \quad , \tag{7.4.5}$$

a quantity called transmission rate. Some confusion is present in textbooks though certainly not in engineering practice, by choosing the time unit so as to make $v = 1$ for convenience, and not stating this very explicitly. Yet the distinction between "uncertainty removed" and "uncertainty removed per time unit" is of economic importance. If (though not only if: see Section 7.6 below) transmission causes garbling, in the formal sense of our Section 4.2, the number $K(\pi,\eta)$ of bits per digit decreases.*/ Thus variations of "uncertainty removed" can affect

---

*/ It is easily seen that, in fact. when the channel is noiseless (i.e., η is an identity matrix) then $J(\pi,\eta) = 0$, $K(\pi,\eta) = H(\pi)$. That is, for given π, uncertainty retained is at its minimum, and uncertainty removed reaches its maximum, when the channel is noiseless.

expected benefit because of possible garbling. But another factor affecting expected benefit is the delay in transmission. An accurate but slow transmission may have the same value to the user as an inaccurate but fast one.

By (7.4.5) the transmission rate depends on v, π, and η. If v and η are kept constant but π varies over the set of all probability vectors of order m, the transmission rate will vary, and its maximum is called the capacity of the channel. It depends on v and η (and thus also on m, for η is of order $m \times n$). However, in theoretical discussion v is usually put = 1, making the capacity, denoted by C, depend on η (and thus m) only. In this notation we have, for any v

$$\max_{\pi} K(\pi,\eta)v = C(\eta)\cdot v \text{ bits per time unit.}$$

7.5 Capacity and cost. It can be presumed that the cost of channel increases with v. It is also usually assumed, I think, that channel cost increases with $C(\eta)$. This assumption was used in Section 6.7, where a formula for $C(\eta)$ was given for η binary and $v = 1$. However, it is not too clear why two channels with two different matrices η, η' should require equal costs (of construction, maintenance and operation) whenever $C(\eta) = C(\eta')$. For example, formula (6.7.1) yields approximately (see Figure 4)

$$C\begin{pmatrix} .83 & .17 \\ .17 & .83 \end{pmatrix} = .3 = C\begin{pmatrix} .5 & .5 \\ .1 & 0 \end{pmatrix}.$$

The matrix on the right is exemplified by a channel which transmits every "no" without fault, but transforms a "yes" into a "no" half of the time:

"You will send me a word (through a very unreliable messenger) only if you decide to come." It is not clear why the use of such a channel should equal in cost the use of a somewhat more reliable messenger who mistakes a "yes" for a "no", or conversely, about one time out of six, as in the matrix on the left. I suppose data on such cost questions are at the disposal of the communication industry. As far as I can see, theoretical literature does, in effect, regard all channels with equal capacity as equivalent with respect to cost. It answers, for example, the question: "What is the best code for a channel with a given capacity?" Yet, the user's economic question should be: "What is it for a channel with a given cost?"

7.6 Does informativeness always increase with "information transmitted?" The answer is no. Let φ be any convex function of a non-negative variable. One such function is

$$\varphi_0(x) = x \log x, \quad 0 \le x \le 1, \tag{7.6.1}$$

since $\varphi_0''(x) = \ln 2/x > 0$. The following has been proved:*/

Theorem: If $\eta^{(1)} = [\eta^{(1)}_{zy^{(1)}}]$ and $\eta^{(2)} = [\eta^{(2)}_{zy^{(2)}}]$ are two information matrices then $\eta^{(1)} > \eta^{(2)}$ if and only if, for any convex function φ,

$$\sum_{y^{(1)}} p(y^{(1)}) \sum_z \varphi(\delta_{y^{(1)}z}) \ge \sum_{y^{(2)}} p(y^{(2)}) \sum_z \varphi(\delta_{y^{(2)}z}), \tag{7.6.2}$$

---

*/ See Blackwell and Girshick [1954], part 4 of Theorem 12.2.2; and /DeGroot [1962].

where, as in Section 7.4, $p(y^{(k)})$ and $\delta_{y^{(k)}z}$ are, respectively, the marginal probability of $y^{(k)}$ and the posterior probability of z, given $y^{(k)}$; both depend on the distribution π and $\eta^{(k)}$. Consider now the particular convex function φ_0 defined in (7.6.1). By the definitions of Section 7.4,

$$\sum_y p(y) \sum_z \varphi_0(\delta_{yz}) = J(\pi,\eta) ,$$

where $\eta = [\eta_{zy}]$. It follows from the above theorem that

$$J(\pi,\eta^{(1)}) \geq J(\pi,\eta^{(2)}) \quad \text{if} \quad \eta^{(1)} > \eta^{(2)} ;$$

it also follows that the converse is not true since the theorem requires condition (7.6.2) to hold for all convex functions and not just for φ_0. It further follows, by (7.4.3), that the condition

$$K(\pi,\eta^{(1)}) \geq K(\pi,\eta^{(2)})$$

is necessary but not sufficient for $\eta^{(1)}$ to be more informative than $\eta^{(2)}$. This means that there exist distributions π and benefit functions (fidelity criteria) θ such that an increase in K, the information transmitted, can be consistent with a decrease in the expected benefit.

7.7 Efficient coding, given a fidelity (benefit) function. Let us continue with the notations of Section 7.4. A channel is characterized by speed v, and by a Markov matrix η, which transforms channel inputs z in Z (occurring with probabilities π_z) into channel outputs y in Y. Now, the channel is a processing link intermediary between two others. On the one hand, at its exit, outputs must be decoded; and, as

before, we identify, in the context of communication theory, the results of decoding (decoded messages) with benefit-relevant actions $\underline{a}$ in $\underline{A}$ (where the sets $\underline{A}$ and $\underline{Z}$ are identical), and, hence, identify the decoding transformation with the strategy α. On the other hand, the benefit-relevant events are not the channel inputs but the messages to be sent. These are transformed into channel inputs by a processing called encoding, possibly preceded by storing, as indicated in Section 7.3. Neglect storing for a moment (i.e., assume it to be characterized by identity transformation, zero costs and zero delays) and denote the messages to be sent by $\underline{s}$ in $\underline{S}$, and their probability distribution by σ. Denote by $\underline{w}$ the speed of inflow of these messages. An encoding Markov matrix ϵ (possibly noiseless) transforms $\underline{S}$ into $\underline{Z}$; and clearly σ and ϵ completely determine the distribution π on $\underline{Z}$. To be feasible, an encoding matrix ϵ is conditioned on some costs and delays, as is the decoding matrix α. These costs and delays are presumably increasing with the length of code words, and also with the number of code words (size of "dictionary"). The pair (ϵ,α) is called <u>code</u>.

Given σ and the benefit (fidelity) function θ on $A \times S$, we can express the expected benefit thus, analogous to (3.6.3):

$$B_{\sigma\theta}(\eta,\epsilon,\alpha) = \sum_{s,z,y,a} \theta(a,s)\sigma_s \epsilon_{sz} \eta_{zy} \alpha_{ya} .$$

If the channel is noiseless, its matrix η is an identity matrix, I. Write

$$\max_{\epsilon,\alpha} B_{\sigma\theta}(I,\epsilon,\alpha) \equiv B_{\sigma\theta}^{\max} ,$$

where the maximization on the left side is over all pairs (ε,α). The notation on the right side is justified if it is proved that

$$B_{\sigma\theta}^{\max} \geq B_{\sigma\theta}(\eta,\varepsilon,\alpha) \quad \text{for all } \eta,\varepsilon,\alpha .$$

In this notation, Shannon's "second theorem" generalized for the case of any fidelity criterion*/ rather than the special one of (7.1.1), can be stated thus:

For any positive k, and given σ, θ, η, there exists a code (ε^*,α^*) such that

$$B_{\sigma\theta}^{\max} - B_{\sigma\theta}(\eta,\varepsilon^*,\alpha^*) \leq k ,$$

provided $$H(\sigma)\cdot w < C(\eta)\cdot v .$$

The left side of the upper inequality becomes the "probability of error" p_e in the special case when all errors are assigned equal penalty as in (7.1.1). The theorem is then reduced to its original formulation. (In addition, the speeds v and w are often taken to be equal.)

The code suggested by Shannon [1960] to prove the generalized second theorem is of a particular form, in two respects:

(1) the encoding consists of two steps, first transforming each message to be sent, s, into what may be called "appropriate action under certainty," a_s in A, such that $\theta(a_s,s) = \max_{a\in A} \theta(a,s)$; and then transforming each a_s into a channel input, another element of A;

(2) the second of these two steps is the same for any two channel matrices η,η' with $C(\eta) = C(\eta')$.

---

*/ See Shannon [1960], Jelinek [1968], Pham [1968].

Unless the sequences of messages are very long, the separation into two steps diminishes the expected benefit. For, as to step (1), imagine S to consist of the following four elements.

s_1: stock will fall by \$10 per share

s_2: stock will fall by \$1 per share

s_3: stock will rise by \$1 per share

s_4: stock will rise by \$10 per share

Let the elements of $\underline{A}$ be

a = sell one share short; a' = buy one share.

Then the benefit function is represented by the following matrix (with rows for actions, columns for messages to be encoded):

$$\beta = \begin{pmatrix} 10 & 1 & -1 & -10 \\ -10 & -1 & 1 & 10 \end{pmatrix} .$$

Under certainty, $\underline{a}$ is appropriate to both s_1 and s_2; and a' is appropriate to both s_3 and s_4. But the loss due to channel noise, when input $\underline{a}$ is sent through the channel, and output a' is received (or conversely), is ten times larger if the message to be encoded is s_1 (or s_4) than when it is s_2 (or s_3). It would be more efficient to encode s_1, s_4 by long ("redundant") sequences of symbols, and s_2, s_3 by shorter ones.

As to step (2), consider the two matrices of Section 7.5. They have equal capacities but, again, it would be efficient, in the case of the right-hand matrix to encode the first, but not the second row with redundancy; while such asymmetry is not called for by the matrix on the left.

However, with or without the particular restriction presented by Shannon's double coding, the code (ϵ^*,α^*) may require, for k small, long code words and waiting for long sequences of messages to be sent. As discussed earlier (Sections 7.2, 7.3), long code words cause delays. Long sequences presuppose storing. Therefore, to realize a code (ϵ^*,α^*) for a small k, it is not possible to neglect (as we have done at the beginning of this Section) the storage of messages that must precede their encoding. And this introduces additional delays.

7.6 Demand for communication links. The cost of each processing link (storing, encoding, transmitting, decoding) will depend on the characteristics of its transformation matrix but it may also, in general, vary with its inputs, as in Section 1.1. Thus the expected cost of encoding will depend on probabilities σ_s of the various messages to be encoded; the expected cost of transmitting will depend on the $\sigma_s\epsilon_{sz}$; and that of decoding, on the $\sigma_s\epsilon_{sz}\eta_{zy}$. And similarly with expected delays. This is simplified if, as in Section 6.6, the cost and delay of each processing depends on the transformation characterizing it (ϵ,η,α) but not on the input; and if the same is assumed of delays. The sum of costs of the links is then subtracted from the expected benefit; and the latter is affected by the delays in the several links, especially because of the diminution of expected benefit, caused by the obsolescence of actions (here: decodings), as in Section 5.

However, most of the existing literature lets each link be associated, not / with its costs and delays, but / with characteristics such as channel capacity, length of the code word, and size of the code dictionary. A

question such as the following is asked:

> given the channel capacity, the (expected) word length, and the code size, how large an expected fidelity can be achieved?*/

Answering such a question would not really provide the set of communication systems efficient from the point of view of a given user, characterized by a fidelity function and a probability distribution of messages to be sent. We remarked in Section 7.5 that two channels with equal capacity (and speed) need not have equal cost. As to the length (or more generally, the expected length) of code words, it is due to delays; and these influence expected utility to the user, not by being added to costs but through a complicated effect on expected benefit, especially by making decisions obsolete, as we have just remarked. Expected utility cannot be decomposed additively into expected benefit, channel capacity and (expected) word length; that is, utility is not linear in these quantities. (Similar considerations would apply to the size of code). Yet without such additivity answers to a question like the one just formulated would not provide the set efficient from a given user's point of view (see Appendix I).

In a sense, the set of non-dominated quadruples (expected fidelity, channel capacity, expected word length, code size) is the result of a

---

*/ This is the formulation given by Wolfowitz [1961], but generalized in two respects: by introducing a general fidelity criterion instead of an equal penalty for all errors; and by permitting the code words to vary in length, thus presumably increasing coding efficiency. I must acknowledge a great debt to Wolfowitz's clear presentation of the economic problem.

crude "averaging" over all users. Delays, being undesirable for all users, are replaced by what amounts to an additive cost, as a make-do. This gives a rough guidance to the supplier of the communication links in estimating the demand for them. The demand of the individual user (if he is "rational") is rather different, and hence that crude average cannot represent the aggregate demand.

8. MARKET FOR INFORMATION

8.1 Demand for systems and sub-systems. Return now to the general outline of purposive processing chains (and networks, for that matter) that we gave in Sections 1 and 2, with especial regard to information systems. The individual user (meta-decider) can achieve a given sequence of transformations only at certain costs and with certain delays (or, more generally, a certain probability distribution of costs and delays). Subject to these constraints, he should maximize the expected benefit simultaneously with respect to all of the transformations. Just like an ideal plant designer decides simultaneously about the size and composition of the personnel as well as of the machine park, the warehouses and the transportation facilities! This is, of course, hardly ever achieved in reality.*/ The humbler meta-decider makes his choices separately for each of several sub-systems; this is what the term "sub-optimization" is often intended to mean, I believe. Hopefully, he partitions the total system in such a way that the complementarity between sub-systems (with regard to expected benefit) is small.

The failure to maximize over all system components simultaneously is just one of many allowances for "lack of rationality" that must be made before we claim a modicum of descriptive validity to the result of aggregating the demands of individual users into the total demand for system components of various kinds, given the constraints.

---

*/ For an attempt to deal more formally with the limitations of the meta-decider ("organizer") see Marschak and Radner [in press], Chapter 9.

8.2 The supply side. The "demand side" of the market, the relation associating the set of constraints with the set of demands, depends on the benefit functions ϑ and the probability distributions π characterizing individual users. The "supply side" is the relation between the constraints and/the supplies, and depends on the "production conditions" ("technology") characterizing each supplier. As usual, the economist is almost completely ignorant of technology.

Let me conclude just with three, rather casual, remarks on these production conditions. It is superfluous to remind the economist that the market is supposed to equalize demand and supply, and the demand and supply constraints.

8.3 Standardization. In many cases, it does not pay to produce "on order." Mass production may be cheaper. This may explain why our Sunday newspaper is so bulky (it gives all things to all subscribers), and why our telephones have such a high fidelity. The individual user is "forced" to purchase information services which, for him, would be wasteful if they were not so cheap.

8.4 Packaging. In our scheme, inquiry was presented as a component separate from storing the data, encoding as separate from transmission, etc. The producer of automata and control mechanisms may find it cheaper to produce them jointly, in fixed "packages." This, again, imposes constraints on the user, similar to those of standardization.

Standardization and packaging are, of course, not peculiar to the production of information services and are present in other markets. I would be grateful for references, especially to writings of a more formal kind.

8.5 Man vs. machine. The competition between machines and human nerves (not muscles) is much discussed today. Some symbol-manipulating services consist in many-to-one mapping, variously called "sorting" and "pattern-recognition." Encoding and decoding are of this nature, but not the (generally noisy) transmission. To be sure, we have, in Section 7, characterized encoding and decoding by Markov matrices, thus allowing for "randomized codes." Such codes have been used for the convenience of mathematical proofs. But, as in any one-person game, there exists an optimal non-randomized choice. Except to allow for (non-rational) error-making encoders and decoders, we may as well consider these activities as many-to-one mappings. In particular, let us consider the "double encoding" proposed by Shannon [1960] and referred to in our Section 7.7. We can imagine the encoder to partition a set of visible or audible stimuli, including verbal sentences, into equivalence classes, variously called "patterns" and "meanings." These are translated, in turn, into the language of channel inputs and outputs, and then decoded back into "patterns" or "meanings." As a special case, we may be little concerned with transmission noise -- newspaper misprints or slips of the tongue. With the channel assumed noiseless the inefficiency of double encoding is removed. The problem of the best code remains: what is the best way to make the receiver (a listener or reader, for example) to "understand" the sender (a lecturer or writer)? The sender must encode into a well-chosen set of patterns, (an "effective style" of speech, or writing, for example), such that the receiver would be able to recognize them, and respond to them by benefit-maximizing actions.

We are told by psycholinguists -- e.g., Miller [1967] -- that man's effectiveness as a channel (and also as a storage facility) is poor compared with inanimate equipment such as telephones (and record tapes). But his coding ability seems superb in many cases. It is variously called "insight," "judgment," "ability to recognize a Gestalt (pattern)"...

APPENDIX I: Requirement of Commensurable Criteria.

In the text, utility was defined on each pair "event, action." It is sometimes useful to introduce an additional concept--the result, r (also called consequence) of the given pair "event, action", and to define utility as a function of the result. The result need not be numerical. For example, the result's values can be "getting cured; dying; continuing in ill health." When the result is a numerical vector, and utility is monotone increasing in each of its components, we call each component a (desirable) criterion.*)

*) In fact, a suggestion has been made to replace the commodity space of usual economic theory by a space of criteria that may "explain" the consumers' preferences: e.g., a car becomes a bundle of criteria such as speed, mileage per gallon of fuel, etc. See Lancaster [1966]. END OF FOOTNOTE

Thus

action = a ; event = z ;

result $r = (r_1,\ldots,r_n)$, with every r_i numerical;

$r_i = \rho_i(a,z)$ (i-th "result function");

utility $u = v(r_1,\ldots,r_n)$;

$v(r_1,\ldots,r_n) > v(r'_1,\ldots,r'_n)$ if

$r_i > r'_i$ for some i, $r_i \geq r'_i$ for all i.

Consider a case when $n = 1$: suppose, e.g., the decision-maker maximizes the expected utility of money profit. The unique component of the criterion vector is then a dollar amount. It is well known that, in this case, expected utility

is not necessarily monotone in expected money profit (independently of some other parameters of the distribution of money profit such as variance) unless utility is linear in money.

Before we generalize to the case of n components, note, as an example, that the pair "minus cost, numerical benefit" constitutes a vector consisting of two criteria. In Section 7.8 the following criteria, used in communication theory, were listed: fidelity criterion; length of code word; size of code; capacity of channel (provided of course that the last three numbers be replaced by their negatives). Given the distribution π of events $\underline{z}$, the action $\underline{a}$ will result in some joint distribution of $r_1,\ldots,r_n$, to be denoted by

$$\pi^a(r_1,\ldots,r_n).$$

Consequently, action $\underline{a}$ will yield expected utility

$$(A.1) \qquad E_a(u) \equiv \sum_{r_1\ldots r_n} v(r_1,\ldots,r_n)\ \pi^a(r_1,\ldots,r_n)\ .$$

Given the action $\underline{a}$, and thus the joint distribution π^a, the marginal probability distribution of a particular criterion, for example of r_1, will be denoted by

$$\pi^a(r_1) \equiv \sum_{r_2\ldots r_n} \pi^a(r_1,\ldots,r_n);$$

no ambiguity results from using the same symbol--here π^a -- for two different functions, made distinguishable by their different arguments, in parentheses.

Then the expected value of r_i, given action $\underline{a}$, is

$$E_a(r_i) \equiv \sum_{r_i} r_i \, \pi^a(r_i) \, . \tag{A.2}$$

The vector of expected criterion values will be denoted by

$$[E_a] = [E_a(r_1),\ldots,E_a(r_n)].$$

Given two actions $\underline{a}$ and $\underline{b}$, we say, as usual, that $[E_a]$ dominates $[E_b]$, and write $[E_a]$ $\underline{\text{dom}}$ $[E_b]$, if

$$E_a(r_i) \geq E_b(r_i) \, , \quad \text{all } i$$

$$E_a(r_i) > E_b(r_i) \, , \quad \text{some } i \, .$$

We shall then also say that action $\underline{a}$ dominates $\underline{b}$ with <u>respect to criterion expectations</u>.

Suppose that

$$\begin{aligned} &E_a(u) > E_b(u) \quad \text{whenever} \\ &[E_a] \ \underline{\text{dom}} \ [E_b] \, . \end{aligned} \tag{A.3}$$

Clearly this is equivalent to saying that expected utility $E_a(u)$ is a monotone increasing function of the expected criterion values $E_a(r_1),\ldots,E_a(r_n)$. If this is the case then, and only then, the feasible action a* (say) that maximizes each of the $E_a(r_i)$ will also maximize $E_a(u)$.

Suppose the utility function v is not known; but condition (A.3), or, equivalently, the monotonicity of $E_a(u)$ with respect to the criterion expectations $E_a(r_1),\ldots,E_a(r_n)$ is known to hold. Then, while it is not possible to determine an optimal action one can at least eliminate all actions that are dominated by some feasible action. The remaining subset of feasible actions will be then, as usual, called the <u>efficient set</u>.

Consider now the case

$$u = v(r) = r_1 + r_2 + \ldots + r_n \ ;$$

then by (A.1),

$$E_a(u) = \sum_{r_1\ldots r_n} r_i\pi^a(r_1,\ldots,r_n)+\ldots+ \sum_{r_1\ldots r_n} r_n\pi^a(r_1,\ldots r_n) \ ;$$

then by (A.2)

$$E_a(u) = \sum_{r_1} r_1\pi^a(r_1)+\ldots+ \sum_{r_n} r_n\pi^a(r_n)$$

$$= E_a(r_1)+\ldots+ E_a(r_n) \ ,$$

an obvious result ("Expectation of sum = sum of expectations"). We shall now prove

<u>Theorem</u>. <u>Expected utility is monotone in expected criterion values if and only if utility is linear in the criteria</u>.

Clearly, the conclusion of this theorem ("the expected utility is monotone in expected criterion values") could be replaced by the following equivalent propositions:

(i) "If action a dominates action b with respect to expected criterion values then a is preferred to b ";

(ii) "The efficient set consists of all those feasible actions which are not dominated, with respect to expected criterion values, by any feasible action."

(iii) "An action that maximizes, over the set of feasible actions, the expected value of each criterion, is optimal."

By substituting any of these three sentences for the conclusion of the Theorem, we obtain three theorems equivalent to it.

The "if" part of Theorem is obvious since a sum is a monotone increasing function of its components. It is unfortunate that the "only if" part is also true. For it follows that unless it is known that utility is additive the computation of expected criterion values loses much of its usefulness: an action b dominated by some other action a with respect to the expected criteria may still be preferable to a , and may indeed by optimal, unless of course some further conditions are known to exist [e.g., distributions $\pi^a(r)$, $\pi^b(r)$,... yielded by all feasible actions are known to belong to some special class-Gaussian, for example].

I shall now give a proof (suggested orally by Roy Radner) of the "only if" part of the above Theorem. Consider three vectors

$$r^o = (r_1^o, \ldots, r_n^o) ,$$

$$r' = (r_1', \ldots, r_n') ,$$

$$\bar{r} = (\bar{r}_1, \ldots, \bar{r}_n) ,$$

where $\bar{r}_i = \alpha r_i^o + (1-\alpha) r_i'$ (all i), and $0 < \alpha < 1$. That is, $\bar{r}$ is a convex combination of r^o and r' (geometrically, $\bar{r}$ is represented by a point on a straight line between r^o and r'). Let two actions, $\underline{a}$ and $\underline{b}$, result, respectively, in the following two joint distributions:

$\pi^a(r)$: $\pi^a(r_1^o, \ldots, r_n^o) = \alpha$, $\pi^a(r_1', \ldots, r_n') = 1-\alpha$;

$\pi^b(r)$: $\pi^b(\bar{r}_1, \ldots, \bar{r}_n) = 1$.

Then for every $i=1, \ldots, n$,

$$E_a(r_i) = \alpha r_i^o + (1-\alpha) r_i' ,$$

$$E_b(r_i) = 1 \cdot \bar{r}_i = \alpha r_i^o + (1-\alpha) r_i' = E_a(r_i) .$$

Hence $E_b(r_i) = E_a(r_i)$, all i .

On the other hand,

$$E_a(u) = \alpha\, v(r^o) + (1-\alpha)\, v(r')$$

$$E_b(u) = v(\alpha\, r^o + (1-\alpha)\, r') .$$

Suppose expected utility of any action is monotone in the expected criterion values (resulting from that action). Then, since $E_b(r_i) = E_a(r_i)$ for all i, we must have

$$E_a(u) = E_b(u),$$

$$\alpha\, v(r^o) + (1-\alpha)\, v(r') = v(\alpha\, r^o + (1-\alpha) r').$$

This is possible only if the function v on the space of vectors $\underline{r}$ is linear, i.e., if there exist $w_i (i=0,1,\ldots,n)$ such that

$$v(r_1,\ldots,r_n) = w_o + \sum^{n} w_i r_i .$$

It is then often said that the criteria are "commensurable" (among each other and with utility itself). A most common case is to convert them in dollars, under the (sometimes tacit) assumption that utility is linear in dollars.

LIST OF REFERENCES

Anderson, N. G. [1969]. Computer interfaced fast analyzers. Science, Vol. 166, pp. 317-324.

Arrow, K. J. and Enthoven, A. C. [1961]. Quasi-concave programming. Econometrica, pp. 779-800.

Ash, R. [1965]. Information Theory. Wiley.

Bellman, R. [1961]. Adaptive Control Processes: A Guided Tour. Princeton University Press.

Blackwell, D. [1953]. Equivalent comparisons of experiments. Annals of Mathematical Statistics, 24, pp. 265-272.

Blackwell, D. and Girshick, A. [1954]. Theory of Games and Statistical Decisions. Wiley.

Breiman, L. [1960]. Another approach to information theory. Electronics Research Laboratory, Series 60, Issue 304. University of California, Berkeley.

Carnap, R. [1950]. Logical Foundations of Probability. University of Chicago Press.

Carnap, R. and Bar-Hillel, Y. [1952]. An outline of a theory of semantic information. Res. Lab. Electronics, Cambridge: M.I.T. Techn. Rept. 247, 1952. Reprinted in Y. Bar-Hillel, Language and Information. Addison-Wesley, 1964.

Carnap, R. [1962]. The aim of inductive logic. Logic, Methodology, and Philosophy of Science, P. Suppes and A. Tarski (eds.). Stanford University Press.

Carnap, R. [1956]. Probability and content measure. Mind, Matter and Method: Essays in Philosophy and Science in Honor of M. Feigl, P. Feyerabend and G. Maxwell (eds.). University of Minnesota Press.

Chernoff, H. [1968]. Optimal stochastic control. Mathematics of the Decision Sciences, Part 2, G. Dantzig and A. F. Veinott (eds.), pp. 149-172. American Mathematical Society.

Cramér, H. [1946]. Mathematical Methods of Statistics. Princeton University Press.

DeGroot, M. H. [1962]. Uncertainty, information and sequential experiments. Annals of Mathematical Statistics, pp. 602-605.

English, J. M., editor [1968]. Cost Effectiveness: Economic Evaluation of Engineering Systems. Wiley.

Good, I. J. [1950]. Probability and the Weighing of Evidence. Hafner.

Good, I. J. [1960]. Weight of evidence, corroboration, explanatory power, information, and the utility of experiments. Journal of the Royal Statistical Society, Series B, pp. 319-331.

Good, I. J. and Toulmin, G. H. [1968]. Coding theorems and weight of evidence. J. Inst. Math. Applics., pp. 94-105.

Hirshleifer, J. [1967]. Notes on the private and social value of information. Working Paper No. 114, Western Management Science Institute, University of California, Los Angeles.

Howard, R. A. [1966]. Information value theory. IEEE Transactions in Systems Science and Cybernetics, Vol. SSC-2, No. 1, pp. 22-34.

Hurwicz, L. [1960]. Optimality and informational efficiency in resource allocation processes. Mathematical Methods in the Social Sciences, K. Arrow et al. (eds.), pp. 27-46. Stanford University Press.

Jellinek, F. [1968]. Probabilistic Information Theory. McGraw-Hill.

Karlin, S. [1959]. Mathematical Methods and Theory in Games, Programming, and Economics. Addison-Wesley.

Koopmans, T. C. [1960]. Stationary ordinal utility and impatience. Econometrica, pp. 287-309.

Lancaster, J. [1966]. Change and innovation in the technology of consumption. American Economic Review.

LaValle, J. [1968]. On cash equivalents and information evaluation in decisions under uncertainty. Journal of American Statistical Association.

Lehmann, E. [1959]. Testing Statistical Hypotheses. Wiley.

Marschak, J. [1954]. Towards an economic theory of information and organization. Decision Processes, R. M. Thrall et al. (eds.), pp. 187-220. Wiley.

Marschak, J. [1960]. Remarks on the economics of information. Contributions to Scientific Management, pp. 79-98. Western Data Processing Center, University of California, Los Angeles.

Marschak, J. [1963]. Adaptive programming. Management Science, pp. 517-526.

Marschak, J. [1964]. Problems in information economics. Management Controls: New Directions in Basic Research, C. P. Bonini et al. (eds.), pp. 38-74. McGraw-Hill.

Marschak, J. [1968A]. Economics of inquiring, communicating, deciding. American Economic Review, Vol. LVIII, No. 2, pp. 1-18.

Marschak, J. [1968B]. Decision-making: Economic aspects. International Encyclopedia of Social Sciences, Vol. 4, pp. 42-55.

Marschak, J. [1970]. The economic man's inductive logic. Volume in Honor of Sir Roy Harrod (forthcoming).

Marschak, J. and Miyasawa, K. [1968]. Economic comparability of information systems. International Economic Review, pp. 137-174.

Marschak, J. and Radner, R. [in press]. Economic Theory of Teams. Cowles Foundation Monograph ..., Yale University Press.

Miller, G. A. [1967]. The Psychology of Communication. Basic Books.

Miller, G. A. and Chomsky, N. [1963]. Finitary models of language users. Handbook of Mathematical Psychology, Vol. II, pp. 419-492. Wiley.

Miyasawa, K. [1968]. Information structures in stochastic programming problems. Management Science, pp. 275-291.

Pham-Huu-Tri, H. M. [1968]. Processing and transmitting information, given a payoff function. Working Paper No. 143, Western Management Science Institute, University of California, Los Angeles (Ph.D. dissertation).

Radner, R. [1967]. Equilibre des marches a terme et au comptant en cas d'incertitude. Cahiers d'Econometrie, No. 9, pp. 30-47.

Radner, R. [1968]. Competitive equilibrium under uncertainty. Econometrica, Vol. 36, No. 1, pp. 31-58.

Raiffa, H. [1968]. Decision Analysis. Addison-Wesley.

Rényi, A. [1966]. Statistics based on information theory. Presented at the European Meeting of Statisticians.

Savage, L. J. [1962]. Bayesian statistics. Recent Developments in Information and Decision Processes, R. F. Machol and P. Gray (eds.). Macmillan.

Shannon, C. E. [1948]. The mathematical theory of communication. Bell System Technical Journal (two papers, reproduced in the book of same title, by Shannon and Weaver, University of Illinois Press, 1949).

Shannon, C. E. [1960]. Coding theorems for a discrete source with a fidelity criterion. Information and Decision Processes, R. E. Machol (ed.), pp. 93-126. McGraw-Hill.

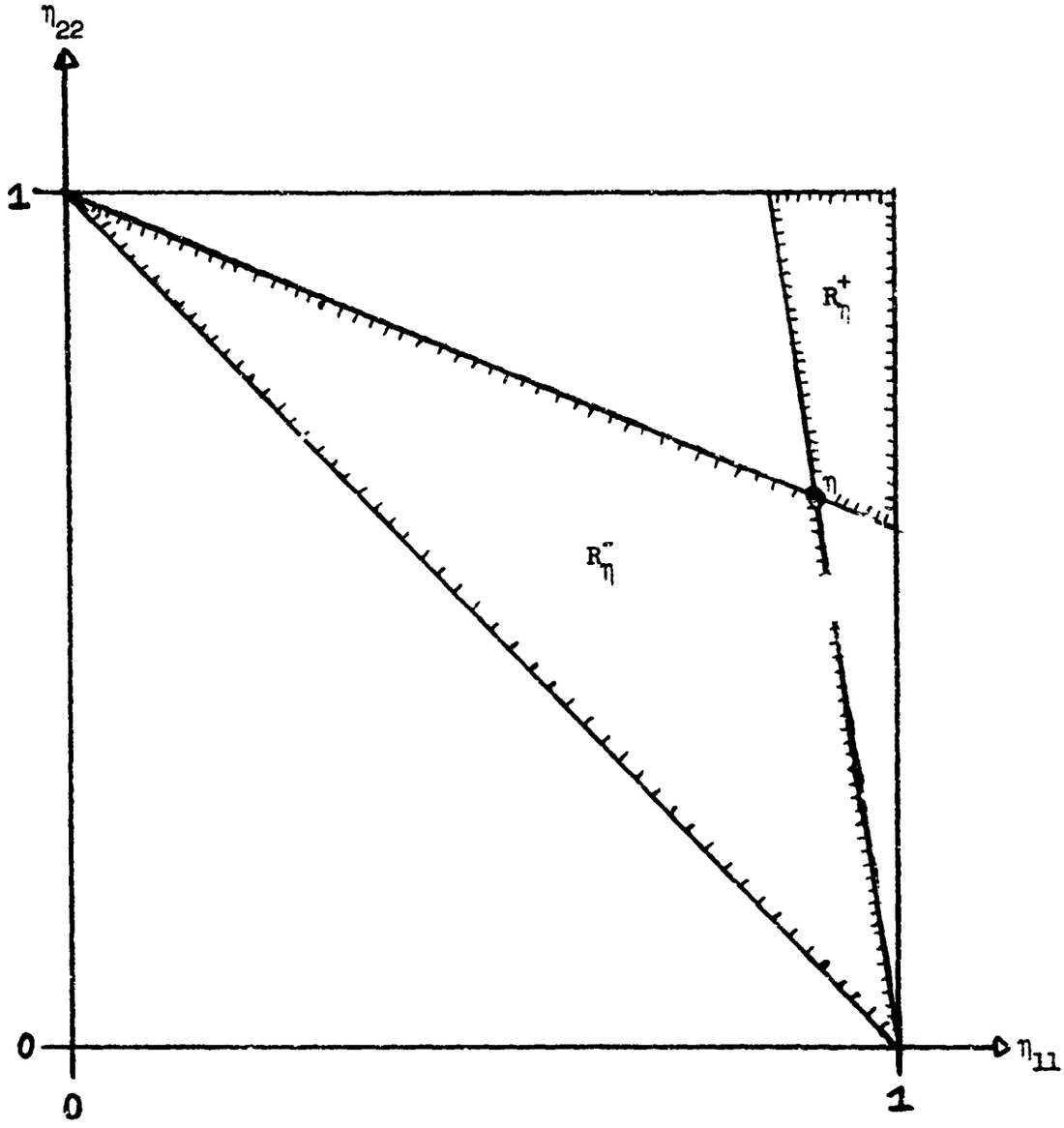
Stigler, G. [1961]. The economics of information. Journal of Political Economy, June, 1961.

Stigler, G. [1962]. Information in the labor market. Journal of Political Economy, October, 1962 (Supplement).

Theil, H. [1967]. Economics and Information Theory. Rand-McNally.

Winter, S. G. [1966]. Binary choice and the supply of memory. Working Papers in Mathematical Economics and Econometrics, No. 97. Berkeley.

Wolfowitz, J. [1961]. Coding Theorems of Information Theory. Springer-Verlag, Berlin.

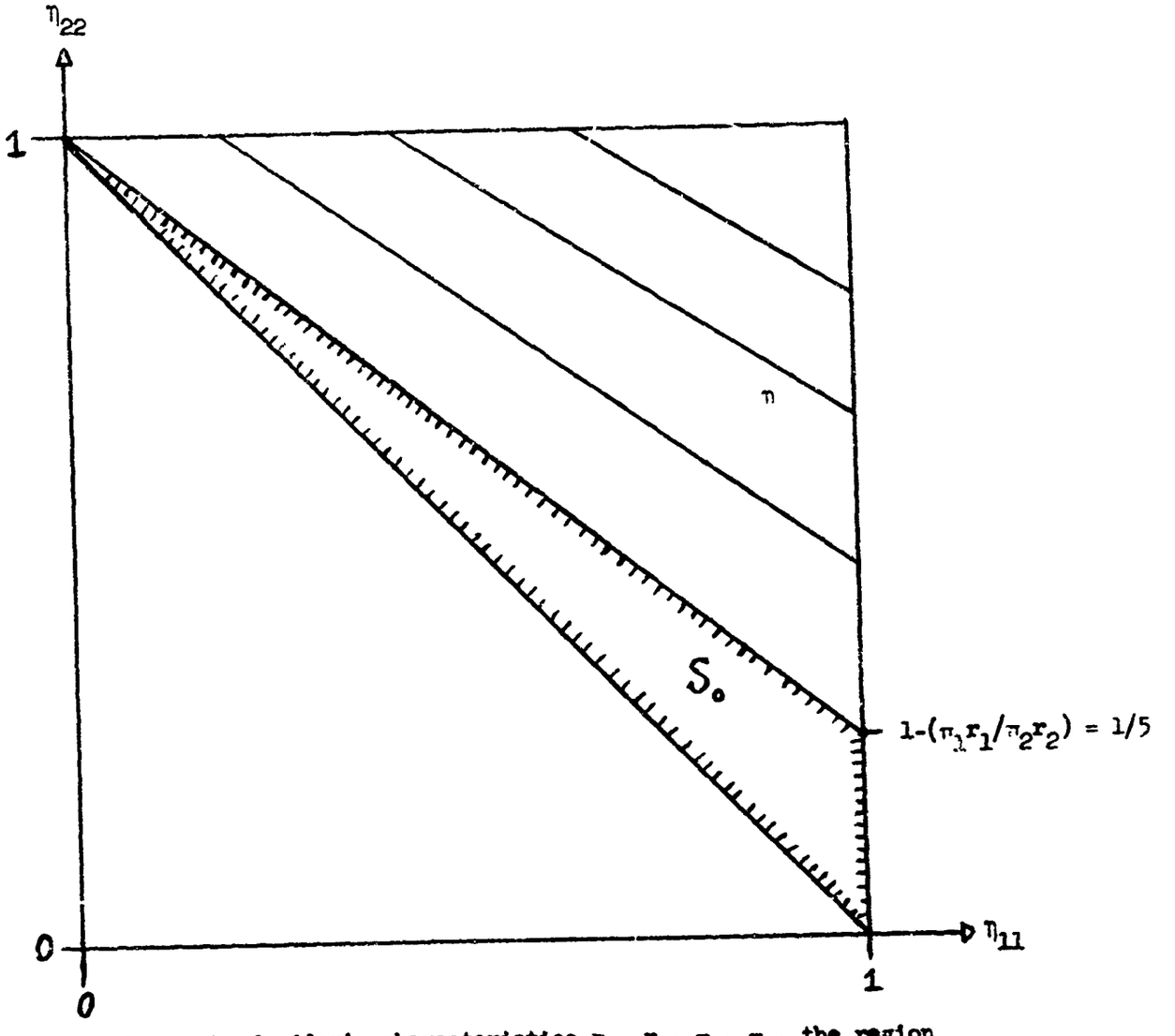


Regions with shaded boundaries are:

R_η^+ consisting of inquiries <u>more informative</u> than η ,

R_η^- consisting of inquiries <u>less informative</u> than η .

FIGURE 1. Informativeness of binary inquiries.



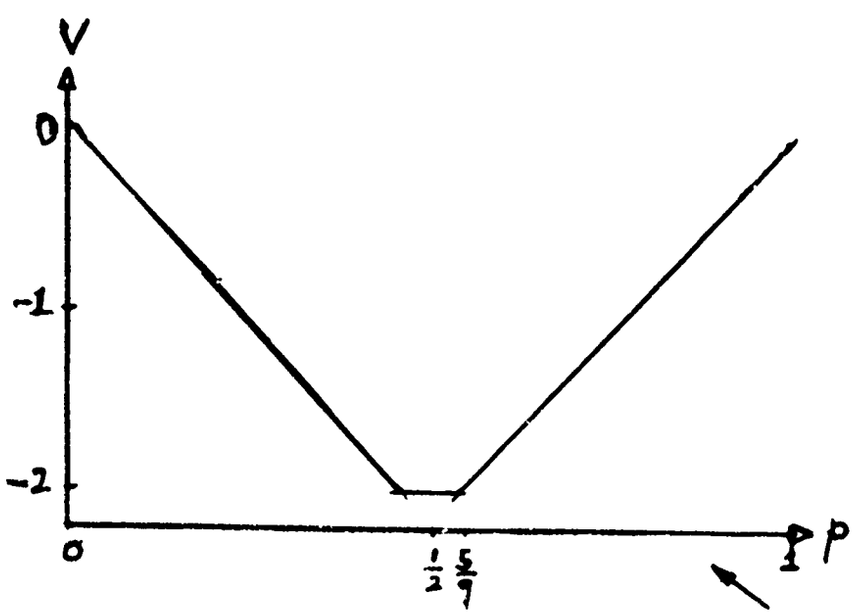
Given the decider's characteristics r_1, r_2, π_1, π_2, the region with shaded boundaries is

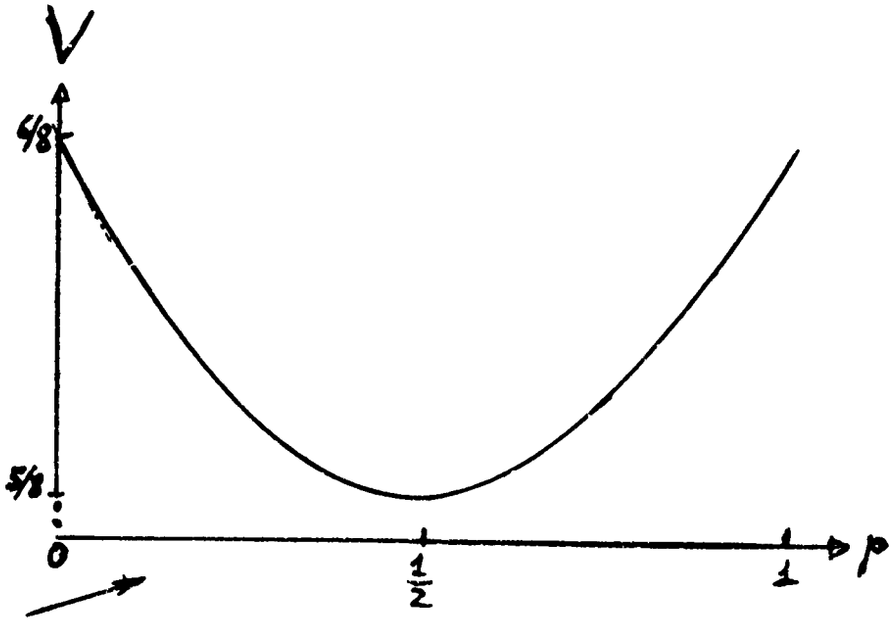
S_0 consisting of <u>useless</u> inquiries.

The remaining Region of the half-square above the main diagonal consists of useful inquiries. It contains indifference lines, all parallel.

η is the same point as on Figure 1.

FIGURE 2. Values of binary inquiries in the case of 2 actions.





Value V of symmetric binary inquiry, when $\pi_1=\pi_2$.

FIGURE 3a

Set of actions A=pair(1,2)

$b_1=b_2= 0$;

$r_1=4$; $r_2=5$

$\beta(a,z)$

	z = 1	2
a= 1	0	-5
2	-4	0

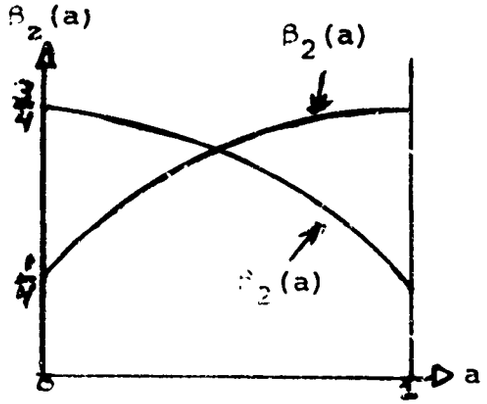
Benefit functions

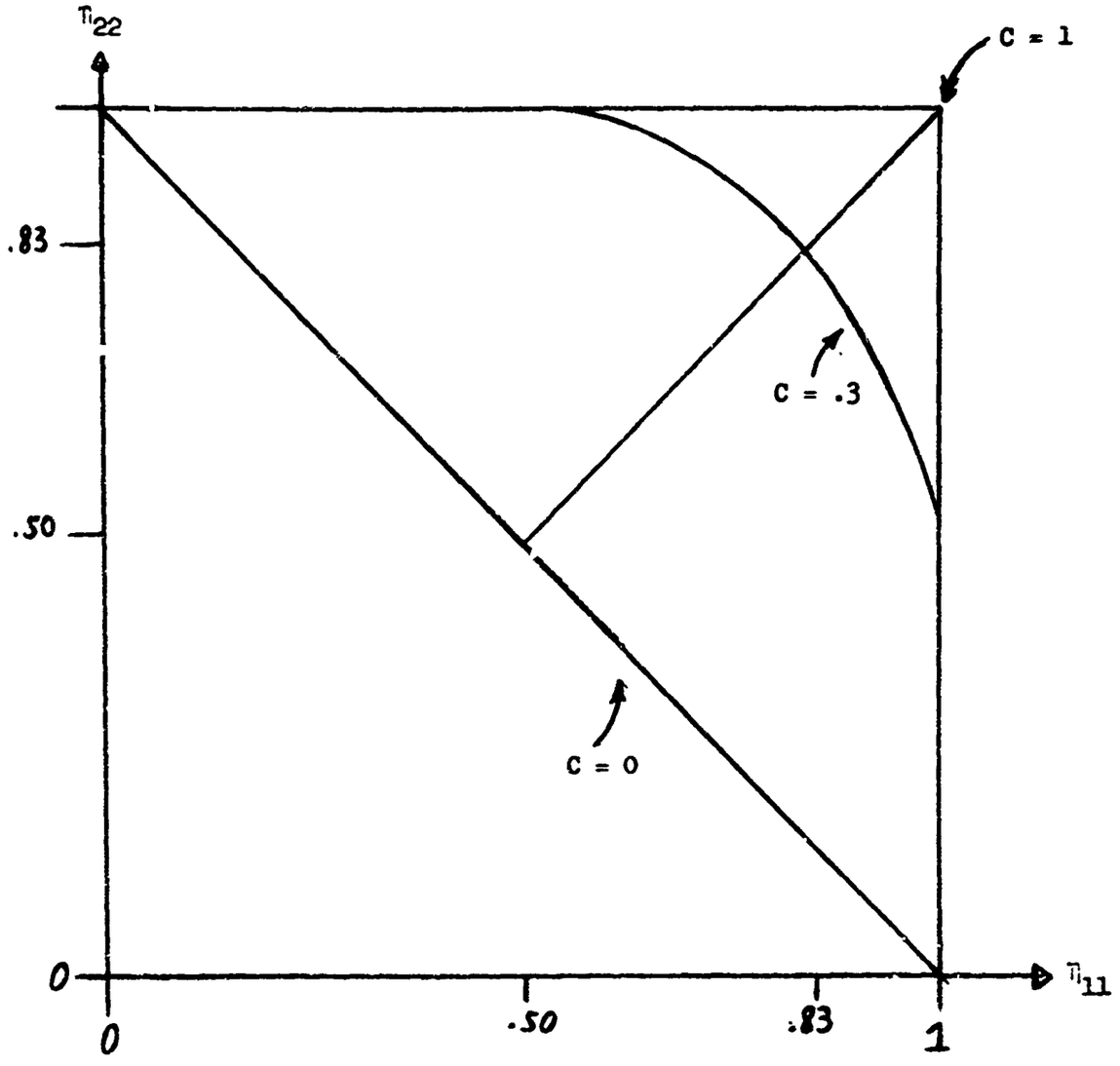
FIGURE 3b

Set of actions A=interval(0,1)

$\beta_1(a)= -\frac{1}{2} a^2+ a + \frac{1}{4}$

$\beta_2(a)= -\frac{1}{2} a^2+ \frac{3}{4}$





Space of binary channel matrices.

Loci of equal channel capacity at unit speed: C = 0, .3, 1.0 .

The upward sloping half-diagonal consists of all binary symmetric channel matrices.

FIGURE 4

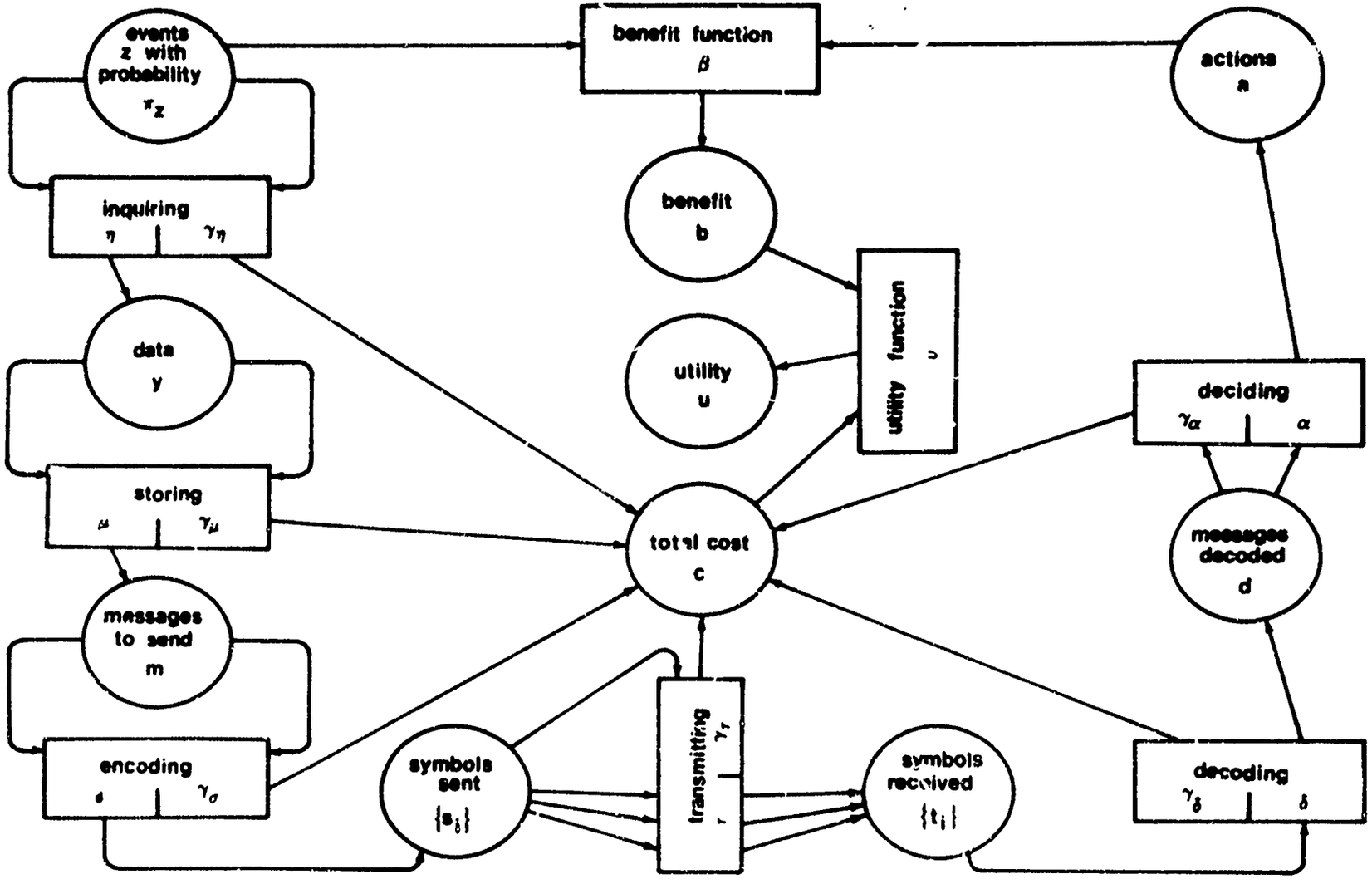


FIGURE 5. Inquiring, Communicating, Deciding.
(Cost functions are $\gamma_\eta, \gamma_\mu, \gamma_\sigma, \gamma_\tau, \gamma_\delta, \gamma_\alpha$.)

users will depend on the joint supply conditions for the various system components. It will thus depend, for example, on the cost economies due to the "packaging" of several components, to standardization and large-scale production. This opens up the question whether social interest is best served by a competitive market in information processing equipment and services, human as well as inanimate.
For simplicity, we have assumed that utility (the quantity whose expected value is maximized by the user) is the difference between costs and benefits. The current literature on communication assumes implicitly that other choice criteria (such as the length of a code word) are additive, and that channels with equal capacity are equally costly. These assumptions may need to be qualified, by studying channel costs and the economic effects of communication delays.

Security Classification

DOCUMENT CONTROL DATA - R & D

Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified

1 ORIGINATING ACTIVITY (Corporate author) Western Management Science Institute	2a REPORT SECURITY CLASSIFICATION Unclassified 2b GROUP

3 REPORT TITLE

ECONOMICS OF INFORMATION SYSTEMS

4 DESCRIPTIVE NOTES (Type of report and inclusive dates)

5 AUTHOR(S) (First name, middle initial, last name)

Jacob Marschak

6 REPORT DATE November, 1969	7a TOTAL NO OF PAGES 7b NO OF REFS
8a CONTRACT OR GRANT NO N00014-69-A-0200-4005, NR 047-041 b. PROJECT NO NSF Grant GS 2041 c. d	9a ORIGINATOR'S REPORT NUMBER(S) Working Paper No. 153 9b OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10 DISTRIBUTION STATEMENT

Distribution is unlimited.

Western Management Science Institute
University of California
Los Angeles, California 90024

11 SUPPLEMENTARY NOTES	12 SPONSORING MILITARY ACTIVITY

13 ABSTRACT

An information system is a chain (or, more generally, a network) of symbol-processing components, each characterized by costs and delays, and by the probabilities of its outputs, given an input. In recent times, statisticians, engineers, and even philosophers have all shown increasing tendency to accept the economist's way of comparing information systems according to their average costs and benefits,--the former depending, in part, on the delays between the events inquired about and the actions decided upon.

Statisticians have concentrated on the economic choice of only these two, the initial and the terminal components of the system: "inquiry" and "decision rule". And they have tended to neglect the processing delays arising in these as well as in the intermediate components of a system. Engineers, on the other hand, have concentrated on the intermediate components that form the "communication sub-chain": "memorizing", "encoding", "transmitting", "decoding". And they have been concerned with the processing delays that depend on the average number of code symbols needed (and thus on the "entropy" to be removed by communication).

The economically minded user must consider the several system components jointly; and it turns out that, in certain important cases, the average difference between the benefit and cost to a user is maximized by large-scale demand. Moreover, the aggregate demand of all

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6801

Security Classification